Interdyscyplinarne Centrum Modelowania Komputerowego
Wydział Matematyczno-Przyrodniczy
Uniwersytet Rzeszowski

METODY STOSOWANIA WIEDZY DZIEDZINOWEJ DO POPRAWIANIA JAKOŚCI KLASYFIKATORÓW

mgr inż. Sylwia Buregwa-Czuma

ROZPRAWA DOKTORSKA

Promotor:
dr hab. Jan G. Bazan, prof. UR
Promotor pomocniczy:
dr Wojciech Rząsa

Rzeszów, 17 maja 2017

Dziękuję wszystkim,
których pomoc przyczyniła się do powstania tej rozprawy,
w szczególności promotorowi dr. hab. Janowi G. Bazanowi, prof. UR
za nieocenioną cierpliwość i pomoc,
doktorowi Wojciechowi Rząsie oraz
profesorowi Andrzejowi Skowronowi,
a także Koleżankom i Kolegom
z Wydziału Matematyczno-Przyrodniczego
Uniwersytetu Rzeszowskiego
za wsparcie i owocną współpracę.

Spis treści

Rozdział 1

Wprowadzenie

Zawartość

1.1 Zakres tematyczny rozprawy

Głównym zadaniem klasyfikacji stanowiącej jedną z ważnych metod eksploracji danych, jest utworzenie modeli, zwanych klasyfikatorami, opisujących zależności pomiędzy zadaną klasą (kategorią) obiektów a ich charakterystyką. Odkryte modele klasyfikacji są następnie wykorzystywane do klasyfikacji nowych obiektów o nieznanej przynależności do klasy (patrz np. [95]). Problem konstrukcji klasyfikatorów często przedstawiany jest jako problem aproksymacji pojęć (klas) na podstawie skończonego zbioru obserwacji zawierającego przykłady pozytywne i negatywne pojęć (patrz np. [94, 18]).

Dane gromadzone w ogromnych ilościach w systemach informatycznych coraz częściej dotyczą złożonych procesów i zjawisk, które nie poddają się klasycznym metodom modelowania. Jednym z ograniczeń istniejących metod jest to, że nie pozwalają one na efektywną aproksymację pojęć złożonych, które mogą być nieostre i wyrażone w języku naturalnym z użyciem różnych innych pojęć występujących w wiedzy dziedzinowej. Przykładami tego rodzaju pojęć są: *zachowanie się pacjenta związane z zagrożeniem życia*, *niebezpieczna jazda samochodem na drodze*, *zachowanie się pacjenta wymagające wykonania odpowiedniego typu plastyki naczyń wieńcowych*, *wystąpienie powikłania po koronarografii*, *nieodwracalna przebudowa oskrzeli jako skutek np. astmy* i inne. Wynika to z faktu, że pojęcia te

znajdują się w zbyt dużej odległości semantycznej od dostępnych danych sensorowych (mierzonych bezpośrednio za pomocą urządzeń czy czujników). Dlatego całkowicie automatyczne podejście do aproksymacji złożonych pojęć za pomocą dostępnych atrybutów (najczęściej są to dane sensorowe) nie prowadzi do klasyfikatorów o zadowalającej jakości (patrz np. [114, 139, 173]).

W literaturze pojawiły się propozycje integracji procesu eksploracji danych z wiedzą dziedzinową (patrz np. [20, 48, 83]) mające umożliwić odkrywanie zależności między pojęciami na różnych poziomach ogólności. Podejście takie ma naśladować proces uczenia się człowieka, w którym wykorzystuje on wcześniej zdobytą wiedzę na temat dotychczasowych zależności między pojęciami [167]. Jednym z wielu wyjaśnień znaczenia dotychczasowej wiedzy w uczeniu się człowieka jest fakt, że wiedza kieruje uwagę w stronę pewnych cech kosztem innych czy też pozwala tworzyć nowe cechy z danych (patrz np. [77]). W rozprawie zaproponowano kilka metod stosowania wiedzy dziedzinowej do poprawiania jakości klasyfikatorów na rożnych etapach procesu budowy modelu.

Z informatycznego punktu widzenia główny problem rozprawy dotyczy zatem budowy klasyfikatorów aproksymujących wybrane, złożone pojęcia z obszaru medycyny. Natomiast z medycznego punktu widzenia, główny problem dotyczy rozpoznawania istotnych zwężeń (stenoz) tętnic wieńcowych w chorobie niedokrwiennej serca (CNS) i potrzeby zabiegu udrażniania naczyń (rewaskularyzacji) przywracającego prawidłowe ukrwienie mięśnia sercowego w oparciu o dane kliniczne oraz wynik badania Holtera (24-godzinny zapis EKG).

Przynależność obiektu (pacjenta) do pojęcia opiera się na wyniku badania angiograficznego tętnic wieńcowych (koronarografii), na podstawie którego wyróżnia się chorobę 1-naczyniową (gdy zwężenie dotyczy tylko jednej tętnicy wieńcowej), 2-naczyniową, 3-naczyniową, 4-naczyniową lub stan bez istotnie zwężonych naczyń. Taka anatomiczna stratyfikacja CNS dostarcza użytecznych wskazówek prognostycznych i jest wykorzystywana do selekcji pacjentów do zabiegu rewaskularyzacji. Pacjenci bez istotnych zwężeń, a więc o najmniejszym nasileniu choroby, generalnie leczeni są zachowawczo, natomiast obecność stenoz wymaga zwykle zabiegu udrażniania naczyń. Pacjenci z chorobą 1- i 2-naczyniową mają duże szanse na leczenie za pomocą przezskórnej interwencji wieńcowej PCI (ang. *percutaneous coronary intervention*), takiej jak angioplastyka balonowa z protezowaniem (stenty) lub bez. Dla pacjentów z chorobą 3, 4-naczyniową natomiast wymaganym leczeniem może być zabieg kardiochirurgiczny, taki jak pomostowanie tętnic wieńcowych CABG (ang. *coronary artery bypass graft*).

Opracowano nowe metody klasyfikacji, a następnie poddano je weryfikacji z użyciem rzeczywistych danych klinicznych dotyczących leczenia pacjentów ze stabilną chorobą niedokrwienną serca, pozyskanych z II Katedry Chorób Wewnętrznych Collegium Medicum Uniwersytetu Jagiellońskiego oraz ogólnodostępnych

zbiorów danych. Wyniki przeprowadzonych doświadczeń wskazują, że są one bardzo obiecujące.

1.2 Motywacja i cel rozprawy

Aproksymacja złożonych pojęć jedynie w oparciu o zbiory danych może napotykać trudności przy konstruowaniu działających efektywnie klasyfikatorów dla rzeczywistych problemów. W związku z tym pojawiły się propozycje zastosowania wiedzy dziedzinowej w procesie konstrukcji klasyfikatorów, której zadaniem jest zawężanie przestrzeni poszukiwań i ułatwienie interpretacji wyników. Wiedza ta jest stosowana głównie na etapie przygotowania danych do eliminacji nieistotnych atrybutów, selekcji najbardziej wartościowych cech czy utworzenia nowych cech.

W literaturze pojawiły się doniesienia, że zastosowanie wiedzy dziedzinowej ma istotny wpływ na wydajność niektórych metod eksploracji danych. Przykładowo w pracach [138] czy [177] badano wpływ wdrożenia wiedzy dziedzinowej na wyniki takich metod klasyfikacji jak: regresja logistyczna, sztuczne sieci neuronowe, metoda k najbliższych sąsiadów k-NN (ang. *k nearest neighbours*), naiwny klasyfikator Bayesa, drzewa decyzyjne oraz metoda wektorów nośnych SVM (ang. *support vector machine*). Poprawa jakości klasyfikacji w porównaniu do modeli bez wiedzy dziedzinowej była najmniejsza dla drzew decyzyjnych oraz dla metody k-NN.

Taki stan rzeczy skłania do postawienia pytania badawczego dotyczącego możliwości efektywnego zastosowania wiedzy dziedzinowej w zakresie klasyfikacji i opisu danych na innych niż dotychczas etapach procesu odkrywania wiedzy.

Za **główny cel** rozprawy postawiono zatem *opracowanie metod wykorzystujących wiedzę dziedzinową do poprawienia jakości klasyfikatorów tworzonych dwiema dobrze znanymi z literatury metodami, tj. metodą drzewa decyzyjnego oraz metodą k najbliższych sąsiadów.*

Główny cel rozprawy był realizowany poprzez następujące **cele szczegółowe** obejmujące:

1. Opracowanie metody ekstrakcji cech opartej na tzw. wzorcach czasowych poprawiającej efektywność klasyfikatorów.

2. Zaproponowanie modyfikacji miary jakości podziału obiektów w węzłach przy generowaniu drzewa decyzyjnego w celu poprawy jakości klasyfikacji za pomocą drzew.

3. Opracowanie metody zwiększania wiarygodności podziałów obiektów w węzłach drzewa decyzyjnego celem poprawy skuteczności klasyfikacji z użyciem drzew.

4. Zdefiniowanie odległości semantycznej pomiędzy obiektami opartej na ontologii pojęć do zwiększenia wydajności klasyfikacji metodą k-NN.

5. Zaproponowanie opisu wpływu czynnika modyfikującego percepcję testowanych obiektów w oparciu o modele klasyfikacji.

Główna teza rozprawy brzmi: *Za pomocą wiedzy dziedzinowej można znacząco polepszyć jakość działania klasyfikatorów modelowanych za pomocą drzew decyzyjnych oraz metodą k najbliższych sąsiadów.* Tezę tę można uszczegółowić za pomocą następujących trzech **tez pomocniczych**.

1. *Proponowane w rozprawie metody konstrukcji klasyfikatorów wykorzystują nowe cechy definiowane przez eksperta, modyfikację jakości podziałów obiektów w węzłach drzewa, cięcia weryfikujące podziały oraz odległość semantyczną pomiędzy obiektami.*

2. *Nowe metody mogą być z powodzeniem stosowane do rozwiązywania rzeczywistych problemów, takich jak nieinwazyjne przewidywanie obecności istotnych zwężeń tętnic wieńcowych wymagających udrożnienia na podstawie informacji klinicznych oraz zapisu EKG metodą Holtera (bez konieczności wykonywania inwazyjnej koronarografii).*

3. *Klasyfikatory tworzone w oparciu o proponowane w rozprawie metody są bardziej powiązane z wiedzą dziedzinową niż modele pozyskane w oparciu o automatyczną analizę zbiorów danych i przez to lepiej uzasadnione.*

Do rozwiązania postawionego problemu badawczego wykorzystano następujące metody, techniki i narzędzia badawcze:

- Analiza i ocena przedmiotu badań;
- Pozyskanie i wstępne opracowanie zbiorów danych zawierających rzeczywiste dane medyczne pacjentów ze stabilną chorobą niedokrwienną serca;
- Modelowanie i implementacja modeli w programie komputerowym;
- Określenie kryteriów oceny rozwiązań;
- Empiryczne zweryfikowanie efektywności zaproponowanych metod dla pozyskanych medycznych zbiorów danych;
- Testowanie opracowanych metod na znanych zbiorach danych, powszechnie stosowanych do oceny metod eksploracji danych [80, 158];
- Analiza porównawcza z innymi metodami;
- Opracowanie wyników badań i postawienie wniosków końcowych.

Główne wyniki rozprawy można podzielić na dwie kategorie: opracowanie metod stosowania wiedzy dziedzinowej do poprawy jakości klasyfikatorów oraz rozwiązanie konkretnych problemów związanych z zadaniem predykcji. Metody służące realizacji celów rozprawy zostały opisane w rozdziałach 4 – 8. Zaproponowane podejścia zostały zaimplementowane i wykorzystane do analizy rzeczywistych zbiorów danych. W przykładowym problemie rozpoznawania obecności istotnych zwężeń naczyń krwionośnych serca, za pomocą pierwszej opracowanej w rozprawie metody utworzono nowe cechy w oparciu o dane temporalne. Cechy te charakteryzują się dużym stopniem przewidywania klas pacjentów, co wykazały przeprowadzone eksperymenty. Druga metoda, stanowiąca propozycję modyfikacji miary jakości podziałów obiektów w węzłach drzewa decyzyjnego, daje także wysoką jakość klasyfikatorów. Kolejno zaproponowano podejście do wyznaczania podziałów węzłów drzewa decyzyjnego z użyciem dodatkowych cięć, nazywanych weryfikującymi. Cięcia weryfikujące realizują ideę ekspertów dziedzinowych, zwiększając pewność podziałów na poszczególnych etapach budowy drzewa. Zastosowanie tej metody daje najlepszą dokładność klasyfikacji spośród wszystkich zaproponowanych metod, co potwierdzają eksperymenty, nie tylko z danymi medycznymi, ale także ze zbiorami danych powszechnie stosowanymi do testowania różnorodnych algorytmów eksploracji danych. W rozprawie podjęto także próbę zdefiniowania odległości semantycznej między obiektami. Do jej wyznaczenia zaprojektowano i utworzono ontologię pojęć dotyczącą głównego medycznego problemu decyzyjnego. Odległości między obiektami mogą być wyznaczone na wiele sposobów, np. z wykorzystaniem odległości Euklidesa czy Manhattan. Jednak odległość semantyczna, w przeciwieństwie do wymienionych, uwzględnia zależności między pojęciami różnych poziomów ontologii, do których należą obiekty. Dało to zdecydowanie lepsze efekty niż zastosowanie odległości wyznaczanych tylko na podstawie danych rejestrowanych przez czujniki. W rozprawie zaproponowana została także metoda opisywania wpływu pewnych czynników modyfikujących postrzeganie obiektów. W przeprowadzonych eksperymentach czynnikiem modyfikującym był wybrany lek, którego zastosowanie zmieniało percepcję zwężeń w tętnicach wieńcowych. Za pomocą tej metody wyznaczono sposób opisywania wpływu wybranej farmakoterapii na postrzeganie zachowania pacjentów.

Praca podejmuje również kilka innych problemów, które pojawiają się w procesach decyzyjnych dotyczących leczenia kardiologicznego. Posłużono się zaproponowanymi metodami poprawiania jakości klasyfikatorów do rozpoznawania, istotnych z punktu widzenia praktycznego, pojęć takich jak "pacjenci ze zdrowym sercem" oraz "obecność zwężeń w dużych tętnicach wieńcowych", uzyskując wysoką rozpoznawalność tych stanów.

1.3 Struktura rozprawy

Rozprawa składa się z dziesięciu rozdziałów. Można w niej wyróżnić trzy części: pierwszą teoretyczną, na którą składają się rozdziały 1, 2 i 3, część drugą, złożoną z rozdziałów od 4 do 8, w której opisuję proponowane metody oraz część trzecią, którą stanowi rozdział 9 poświęcony opisowi eksperymentów i rozdział 10 zawierający podsumowanie oraz najważniejsze kierunki dalszych badań.

W Rozdziale 2 przedstawiono wprowadzenie do tematyki procesu odkrywania wiedzy z danych z uwzględnieniem wiedzy dziedzinowej. Omówiono tu podstawowe pojęcia i zagadnienia związane z wiedzą dziedzinową, podjęto próbę zdefiniowania tego pojęcia oraz przedstawiono sposoby reprezentacji tego typu wiedzy, wykorzystywane w dalszej części rozprawy. Przeprowadzono także analizę literatury związanej z zastosowaniem wiedzy dziedzinowej w procesie odkrywania wiedzy. Rozdział 3 zawiera krótki opis zadania klasyfikacji z użyciem wybranych do badań metod, tj. drzew decyzyjnych oraz metody k najbliższych sąsiadów wraz z przedstawieniem sposobów oceny efektywności tych klasyfikatorów.

Pierwszą z proponowanych metod, polegającą na wyznaczaniu wzorców czasowych wykorzystywanych do aproksymacji złożonych pojęć, przedstawiono w Rozdziale 4. Do definiowania cech odpowiednich dla okien czasowych wykorzystano eksperta, który w oparciu o wiedzę dziedzinową proponuje nie tylko same cechy, ale także sposób wyznaczania ich wartości w poszczególnych oknach czasowych. W Rozdziale 5 zdefiniowano drugą z metod mającą na celu modyfikację oceny jakości podziału obiektów w węzłach drzewa decyzyjnego. Metoda wykorzystuje informacje na temat zróżnicowania wewnątrz klas decyzyjnych do wykrywania subtelnych różnic między przykładami pozytywnymi i negatywnymi aproksymowanego pojęcia. Trzecia metoda oparta na zastosowaniu cięć weryfikujących realizujących ideę ekspertów dziedzinowych opisana została w Rozdziale 6. Wykorzystuje ona dodatkową wiedzę dziedzinową zawartą w zbiorach danych do zwiększania wiarygodności podziałów węzłów drzewa decyzyjnego. Rozdział 7 zawiera definicję odległości semantycznej pomiędzy obiektami, opartą na ontologii pojęć, mającą na celu zwiększenie wydajności klasyfikacji metodą k-NN. Do wyznaczenia odległości opracowano ontologię pojęć dla rzeczywistego problemu, tj. choroby niedokrwiennej serca. Rozdział 8 przestawia propozycję opisywania wpływu czynnika modyfikującego percepcję obiektów w oparciu o modele klasyfikacji. Metoda ta oparta jest na drzewie decyzyjnym, w którym jako kryterium wyboru najlepszego podziału zaproponowano miarę opartą na odległości pomiędzy grupami obiektów, wyliczaną z wykorzystaniem teorii prawdopodobieństwa i metod statystycznych.

Rozdział 9 zawiera opis przeprowadzonych badań eksperymentalnych, charakterystykę danych użytych do testów oraz wyniki eksperymentów mających na celu sprawdzenie efektywności zaproponowanych metod. W rozdziale tym dokonano także uporządkowania wyników oraz przedstawiono zestawienie najważniejszych

wyników służących do postawienia wniosków końcowych.

Podsumowanie, wnioski oraz kierunki dalszych badań zamieszczono w Rozdziale 10.

Rozdział 2

Wiedza dziedzinowa w procesie odkrywania wiedzy z danych

Zawartość

2.1 Proces odkrywania wiedzy z danych

W wyniku intensywnego rozwoju technologii generowania, gromadzenia i przetwarzania danych towarzyszących upowszechnieniu systemów informatycznych, ludzkość dysponuje coraz większymi zbiorami danych. Możliwości analizowania i rozu-

mienia tak dużych zbiorów danych są ograniczone i tylko niewielka ich część jest analizowana i wykorzystywana w praktyce.

Dążenie do efektywnego i racjonalnego wykorzystania nagromadzonej w tych danych wiedzy, przyczyniło się do rozwoju metod i technologii *eksploracji danych*. Eksploracja danych, określana alternatywnie jako: *ekstrakcja wiedzy, drążenie danych, inteligencja biznesowa, pozyskiwanie wiedzy* (ang. *Data Mining, DM)*, wprowadza nową jakość i zakres analiz danych. Zgodnie z definicją przedstawioną w [64], eksploracja danych jest "analizą (często ogromnych) zbiorów danych obserwacyjnych celem znalezienia nieoczekiwanych związków oraz podsumowania danych na oryginalne sposoby, które są zarówno zrozumiałe, jak i przydatne dla ich właściciela". Eksploracja danych stanowi dziedzinę informatyki integrującą szereg dyscyplin badawczych, takich jak m.in.: statystyka, teoria informacji, modelowanie matematyczne, sztuczna inteligencja, systemy baz danych i hurtownie danych, obliczenia równoległe czy optymalizacja i wizualizacja obliczeń. Wykorzystuje również techniki i metody opracowane na gruncie systemów wyszukiwania informacji, rozpoznawania obrazów, analizy danych przestrzennych, przetwarzania sygnałów, grafiki komputerowej, technologii internetowych czy bioinformatyki. Eksploracja danych znalazła zastosowanie praktycznie w każdej dziedzinie życia, takich jak: nauka, medycyna, przemysł, handel i marketing, administracja, finanse i bankowość czy telekomunikacja [85, 129, 84, 71, 178].

Eksploracja danych często umiejscawiana jest w szerszym kontekście procesu odkrywania wiedzy w bazach danych, określanego jako KDD (ang. *Knowledge Discovery in Databases* czy *database mining*). Zadaniem KDD jest odkrywanie nietrywialnych, dotychczas nieznanych zależności, potencjalnie użytecznych reguł, związków, podobieństw czy trendów, ogólnie nazywanych wzorcami (ang. *patterns*) [53]. Odkrywane wzorce mają najczęściej postać reguł logicznych, klasyfikatorów (np. drzew decyzyjnych), zbiorów skupień czy wykresów. Termin *odkrywanie wiedzy* ma ogólniejszy charakter niż eksploracja danych i dotyczy całego procesu odkrywania wiedzy, stanowiącego zbiór kroków przekształcających surowe dane w zbiór wzorców, które mogą być wykorzystane we wspomaganiu podejmowania decyzji. Proces KDD obejmuje takie etapy jak: wybór danych do badania, wstępna obróbka danych, ich transformacja, eksploracja danych oraz interpretacja i ocena odkrytych struktur [64, 40].

Konwencjonalne metody odkrywania wiedzy napotykają jednak poważne trudności w rozwiązywaniu problemów decyzyjnych dotyczących rzeczywistych zagadnień. Wśród przyczyn tego problemu można wymienić dużą złożoność pojęć, których te problemy dotyczą, nieadekwatną reprezentację przypadków reprezentujących pojęcia, zaszumienie danych lub ich niekompletność. Prowadzić to może do odkrywania zbyt wielu reguł, których analiza jest czasochłonna i często niemożliwa do wykonania w rozsądnym czasie, dużej złożoności obliczeniowej czy zjawiska

zwanego przeuczeniem (ang. *overfitting*). Bardzo duże rozmiary baz danych czynią proces odkrywania kosztownym obliczeniowo. Dużym wyzwaniem jest także pozyskiwanie danych czy reprezentacja wiedzy.

Ogrom danych oraz przedstawione trudności w odkrywaniu wiedzy zmuszają do korzystania z podejść, które ograniczają przestrzeń poszukiwań czy skupiają się na wybranej części odkrytych wzorców. Jednym z rozwiązań może być zastosowanie dodatkowej wiedzy, zwanej *wiedzą dziedzinową WD* (ang. *domain knowledge, background knowledge*). Jednym ze źródeł tej wiedzy jest wiedza ekspertów w danej dziedzinie. Wiedza ta umożliwia m.in. zmniejszenie przestrzeni do przeszukania przy szukaniu wzorców.

W rozprawie podjęto próbę odpowiedzi na pytanie, czy zastosowanie wiedzy dziedzinowej w zadaniu klasyfikacji może poprawiać jej efektywność i jak można to robić.

2.2 Przesłanki dla zastosowania wiedzy dziedzinowej do poprawienia klasyfikatorów

Metody budowy klasyfikatorów oparte na tablicach decyzyjnych często napotykają trudności związane z konstrukcją takiej tablicy, która umożliwi budowę efektywnego klasyfikatora. Przyczynami tego zjawiska mogą być trudności w zdefiniowaniu odpowiednich cech aproksymujących dane pojęcie (problem ekstrakcji cech) lub problemy z doborem właściwych cech spośród dostępnych w zbiorze danych (problem selekcji cech). Ponadto mogą pojawiać się problemy z przypisaniem obiektu do danego pojęcia, zwłaszcza gdy pojęcie jest opisane w złożony sposób i wyrażone w języku naturalnym lub wystarczająco dokładną aproksymacją takiego pojęcia za pomocą dostępnych cech w sytuacji, gdy atrybuty warunkowe posiadają bardzo dużo wartości przy jednocześnie małej liczbie obiektów treningowych. Dużym problemem jest także określenie miary podobieństwa obiektów w kontekście wartości atrybutu decyzyjnego, na przykład w przypadku, gdy wartość atrybutu decyzyjnego jest złożona, np. ma postać grafu zachowania, planu czy algorytmu wykonania zadania (patrz [20]).

Często efektem powyższych problemów jest to, że wiele z klasycznych metod tworzenia klasyfikatorów działając w oparciu o ustalone heurystyki selekcji czy dyskretyzacji nie prowadzi do zadowalających efektów w zakresie konstrukcji klasyfikatorów dla danego problemu decyzyjnego.

Powyższe trudności pojawiają się szczególnie w przypadku potrzeby aproksymacji tzw. złożonych pojęć czasowo-przestrzennych. Są to pojęcia wyrażone w języku naturalnym na dużo wyższym poziomie abstrakcji niż tzw. *dane sensorowe*, stosowane do tej pory najczęściej do aproksymowania pojęć. Przykładami takich

pojęć są: *zachowanie się pacjenta związane z zagrożeniem życia*, *zachowanie się pacjenta wymagające wykonania odpowiedniego typu plastyki naczyń wieńcowych*, *wystąpienie powikłania po koronarografii*, *nieodwracalna przebudowa oskrzeli jako skutek np. astmy*, *bezpieczna jazda samochodem* itd. Za dane sensorowe będą rozumiane tutaj dane pochodzące z czujników, wchodzących w skład różnego rodzaju systemów monitorowania procesów czy stanu obiektów, dotyczące pomiarów pojedynczych parametrów.

Istotnym ograniczeniem istniejących metod jest między innymi fakt, że do efektywnej aproksymacji złożonych pojęć potrzebne jest odkrycie niezwykle złożonych wzorców. Intuicyjnie rzecz biorąc, takie pojęcia są zbyt oddalone w sensie semantycznym od dostępnych pojęć, np. reprezentowanych za pomocą sensorów. W konsekwencji przestrzeń poszukiwań, którą należy przeszukać celem odnalezienia wzorców istotnych dla aproksymacji jest tak duża, że jej eksploracja jest niemożliwa do realizacji przy użyciu istniejących metod oraz technologii. Jak się okazuje, uzyskanie wysokiej jakości aproksymacji złożonych pojęć z dostępnych pojęć, zwykle zdefiniowanych dla danych sensorowych, w całkowicie automatyczny sposób za pomocą istniejących systemów stanowi ogromny problem, ponieważ otrzymywane klasyfikatory posiadają niesatysfakcjonującą jakość (patrz [20, 173, 114]).

Ostatnio w literaturze [48, 173] wskazuje się, że jednym z wyzwań eksploracji danych jest odkrycie metod łączących wykrywanie wzorców i pojęć z wiedzą dziedzinową. Wiedza ta dotyczy pojęć występujących w danej dziedzinie oraz różnorodnych związków pomiędzy tymi pojęciami i znacznie przekracza wiedzę zebraną w zbiorach danych. Zwykle jest reprezentowana w języku naturalnym i pozyskiwana poprzez dialog ze specjalistą w danej dziedzinie.

Ogólna motywacja stosowania wiedzy dziedzinowej do polepszenia jakości klasyfikatorów jest taka, że wiedza dziedzinowa może być pomocna w wyborze właściwego dla danego zbioru danych modelu klasyfikatora przy wykorzystaniu określonego paradygmatu tworzenia klasyfikatora (np. reguły decyzyjne, drzewa decyzyjne, metody statystyczne itd.). Przestrzeń możliwych klasyfikatorów przy wykorzystaniu określonego paradygmatu tworzenia klasyfikatora może być bardzo duża. Tymczasem na potrzeby praktycznych zastosowań, konieczny jest wybór tylko jednego lub kilku klasyfikatorów, które będą możliwie najlepiej działać i to nie tylko dla danych treningowych, ale także testowych. Każda z klasycznych metod zwykle oparta jest na jakiejś heurystyce, która dostarcza określonego klasyfikatora. Jeśli heurystyki te nie uwzględniają w wystarczającym stopniu wiedzy dziedzinowej na temat rozpatrywanych problemów, może się zdarzyć, że skonstruowane klasyfikatory, choć dobrze dopasowane do danych treningowych, są mało efektywne dla danych testowych. Klasycznym przykładem jest tutaj sytuacja, gdy metoda tworzenia klasyfikatora preferuje pewien atrybut numeryczny, który na próbce treningowej doskonale dyskryminuje klasy decyzyjne pewnego diagnostycznego problemu

medycznego, ale o atrybucie tym eksperci wiedzą, że ma bardzo niewielkie znaczenie diagnostyczne i klasyfikator nie powinien uwzględniać tej cechy. Przykład ten pokazuje zatem, że często warto jest użyć wiedzy dziedzinowej do wybrania lepszego klasyfikatora.

Przesłanką do zastosowania wiedzy dziedzinowej do konstrukcji klasyfikatorów jest także fakt, że w praktyce często posiadamy stosunkowo niewielkie zbiory danych, które w sensie statystycznym nie są reprezentatywne w stosunku do rozpatrywanych problemów decyzyjnych. W takich przypadkach, zastosowanie dodatkowej wiedzy dziedzinowej wydaje się być jedynym sposobem uzyskania efektywnych w praktyce klasyfikatorów.

Jednym ze sposobów użycia wiedzy dziedzinowej do polepszenia klasyfikatorów jest użycie jej bezpośrednio do poprawy efektywności istniejących podejść wykorzystujących tablicę decyzyjną. Takie podejście było już od dawna praktykowane. Na przykład, przy generowaniu reguł decyzyjnych można wprowadzać podpowiedziane przez eksperta wagi klas decyzyjnych, które mogą być użyte w metodzie konstrukcji klasyfikatora lub w metodzie klasyfikacji nowych przypadków. Jeśli zatem budujemy klasyfikator regułowy do rozpoznawania jakiejś choroby (dwie klasy decyzyjne: "chory" - pacjent choruje na daną chorobę, "zdrowy" - pacjent nie choruje na daną chorobę), to zwiększenie wagi klasy decyzyjnej "chory" często pozwala na zmniejszenie liczby fałszywie zaklasyfikowanych pacjentów jako "zdrowy". Ma to znaczenie dla zwiększenia tzw. specyficzności klasyfikacji. Natomiast wprowadzenie wag atrybutów może, dla przykładu, pomóc przy wybraniu odpowiedniego reduktu z wyznaczonego zbioru reduktów (redukt to minimalny zbiór atrybutów zachowujących rozróżnialność obiektów tak jak wszystkie atrybuty), który ma być wykorzystany do dalszej konstrukcji klasyfikatora [18].

Innym przykładem polepszenia jakości klasyfikatora jest dyskretyzacja atrybutów wsparta za pomocą wiedzy dziedzinowej.

W rozprawie rozpatrywana jest dyskretyzacja z nadzorem, tzn. chodzi o takie metody dyskretyzacji, które używają do swojego działania wartości atrybutu decyzyjnego dla przypadków treningowych. Istnieje wiele metod dyskretyzacji z nadzorem, które oparte są na różnych heurystykach. W rozprawie stosowane jest podejście oparte na tworzeniu tzw. drzewa decyzyjnego lokalnej dyskretyzacji (patrz np. [14]). Jest to drzewo binarne, tworzone za pomocą wielokrotnych podziałów danego zbioru na dwie grupy obiektów za pomocą wartości wybranych atrybutów. Sposób wybrania atrybutu oraz jego wartości (dla atrybutów numerycznych często zwanej cięciem), wykorzystywanych do podziału jest kluczowym elementem omawianej metody budowy drzewa lokalnej dyskretyzacji i powinien wiązać się z analizą wartości atrybutu decyzyjnego dla obiektów treningowych. Jako miarę jakości cięcia, można wykorzystać np. liczbę par obiektów rozróżnianych przez cięcie i mających różne wartości atrybutu decyzyjnego. Jeśli wyznaczymy wartość tej

miary dla wszystkich potencjalnych par (atrybut, wartość), to możemy zachłannie wybrać jedną taką parę i na jej podstawie dokonać podziału całego zbioru obiektów na dwie części. W korzeniu drzewa mamy zatem cały zbiór obiektów. Następnie rekurencyjnie stosujemy tę samą procedurę podziału dla pojawiających się części, które przyporządkowujemy do węzłów drzewa coraz wyższego poziomu. Warunek zakończenia podziału (warunek stopu) jest tak skonstruowany, że dana część nie jest dzielona (zostaje liściem drzewa), gdy należą do niej tylko obiekty z jednej klasy decyzyjnej (ewentualnie obiekty danej klasy stanowią określony procent, który traktujemy jako parametr metody) albo dzielenie nie daje już żadnego efektu (wszystkie potencjalne cięcia nie rozróżniają już par obiektów z różnych klas decyzyjnych).

Po skonstruowaniu takiego drzewa uzyskujemy zestaw cięć, które mogą posłużyć do skonstruowania nowych binarnych atrybutów dla danej tablicy decyzyjnej. Informacje o cięciach można także zgrupować według atrybutów i wyznaczyć nowe wartości atrybutów wejściowej tablicy decyzyjnej. Taką tablicę decyzyjną nazywa się tablicą zdyskretyzowaną, a każdy z atrybutów tej tablicy ma wartości symboliczne wynikające z pierwotnych wartości numerycznych.

Opisany wyżej prosty sposób obliczania miary jakości cięcia może zostać zmodyfikowany za pomocą wprowadzenia wiedzy dziedzinowej. Np. dla problemu rozpoznawania pacjentów, którzy wymagają rewaskularyzacji w oparciu o sygnał EKG uzyskany metodą Holtera, miarę jakości cięć można zmodyfikować poprzez wprowadzenie wiedzy o wewnętrznym zróżnicowaniu klas decyzyjnych. Mianowicie dane cięcie otrzymuje określoną liczbę punktów za każdą parę rozróżnionych pacjentów z różną liczbą zmienionych naczyń, przy czym punkty przydzielane są przez eksperta dziedzinowego (Rozdz. 5, Tab. 5.1).

Zauważmy, że w powyższej metodzie jakość cięć modyfikowana jest przez dodatkową informację o pacjencie (liczba zwężonych naczyń). Jest to możliwe dlatego, że do oryginalnego binarnego atrybutu decyzyjnego (obecność istotnych zwężeń) dokładamy inny, na potrzeby obliczania miary cięć. Analiza danych wykorzystująca opisaną wyżej metodę mierzenia jakości cięć przeprowadzona na danych klinicznych i laboratoryjnych oraz zapisach 24-godzinnego monitorowania EKG metodą Holtera, doprowadziła do opracowania wstępnych metod, które mają czułość 94% (patrz [14]). Wyniki te są dużo lepsze od rezultatów metody z klasyczną miarą jakości cięć.

Powyższe fakty dobrze pokazują, że zastosowanie dodatkowej wiedzy dziedzinowej może spowodować poprawę jakości klasyfikatora, chociaż z całą pewnością nie wyczerpuje wszystkich możliwości w zakresie polepszenia dyskretyzacji za pomocą wiedzy dziedzinowej. Łatwo zauważyć, że jakość cięcia oddzielającego parę obiektów mogłaby być zależna od wartości atrybutu decyzyjnego w bardziej skomplikowany sposób. Na przykład dla pary obiektów, które w jakimś sensie bardziej

różnią się od siebie wartością decyzji, miara cięcia mogłaby mieć zwiększoną wartość w bardziej subtelny sposób. Wymaga to jednak specjalnych metod mierzenia podobieństwa pomiędzy obiektami w kontekście wartości atrybutu decyzyjnego. Szczególnie trudna sytuacja pojawia się wtedy, gdy wartości atrybutu decyzyjnego są w jakimś sensie złożonymi wartościami (np. wektorem wartości, wzorcem zachowania, planem itd.). Nasuwa się wniosek, że w takim przypadku do mierzenia podobieństwa pomiędzy obiektami w kontekście wartości atrybutu decyzyjnego potrzebna jest dodatkowa wiedza dziedzinowa. Przykładową metodą tego typu byłaby metoda oparta na specjalnie skonstruowanej ontologi pojęć, podobna do tej, jakiej użyto do mierzenia podobieństwa pomiędzy planami w [16]. Zauważmy jednak, że prowadzenie badań w tym kierunku wymaga dużego zaangażowania ze strony ekspertów medycznych celem zdefiniowania ontologii medycznej opisującej podobieństwo pomiędzy pacjentami w kontekście potrzeby rewaskularyzacji. Można się spodziewać, że metoda mierzenia podobieństwa oparta na tej ontologii może znacząco polepszyć jakość rozpoznawania potrzeby rewaskularyzacji dla pacjentów testowych.

2.3 Wybrane sposoby reprezentacji danych i wiedzy

Aby wiedza dziedzinowa mogła być zastosowana w procesie odkrywania wiedzy z danych musi być sformalizowana i zaprezentowana w formie jawnej. Sposób w jaki dane, a także wiedza, są usystematyzowane w zbiorach, determinuje możliwości ich efektywnego wykorzystania oraz prowadzenia różnych analiz. Z tego powodu przedstawiane są one w postaci różnych schematów zwanych modelami. Model danych definiowany jest jako zbiór struktur, który służy do opisu i reprezentacji wybranych aspektów świata rzeczywistego w systemach komputerowych [89]. Pod pojęciem reprezentacji wiedzy należny rozumieć sposób odwzorowania wiedzy z pewnej dziedziny za pomocą określonych struktur danych oraz języka reprezentacji wiedzy używanego przez system, który ją przetwarza. Według [28] reprezentowanie wiedzy polega na "tworzeniu opisów świata lub jego stanów". Reprezentacja wiedzy jest pojęciem podstawowym dla procesów decyzyjnych oraz wnioskowania. Głównymi elementami reprezentacji wiedzy są syntaktyka, jako forma reprezentacji (język), semantyka, czyli znaczenie reprezentowanej wiedzy (interpretacja) oraz wnioskowanie, czyli wyprowadzenie wniosków prowadzące do wykorzystania wiedzy.

Wiedza może być zapisana na wiele sposobów, takich jak reprezentacja wykorzystująca język naturalny i zapis w postaci tekstu, diagramów procesów czy reguł. Niestety, komputery nie są w stanie zrozumieć ludzkiej wiedzy bezpośrednio, co wymaga przekładania jej na zrozumiałą dla systemów komputerowych. Zagadnienie to napotyka jednak na szereg problemów, takich jak np. reprezentowanie czasu, idei, przekonań czy informacji niepewnych lub niekompletnych.

Szeroko rozpowszechnionymi i dobrze poznanymi metodami reprezentacji wiedzy są metody symboliczne, wśród których do najczęściej stosowanych zalicza się [157, 98]:

- Metody bazujące na zastosowaniu logiki:
  - Logika konwencjonalna: rachunek zdań, rachunek predykatów;
  - Logika niekonwencjonalna (rozmyta, wielowartościowa);
  - Metody wykorzystujące zapis stwierdzeń;
  - Metody wykorzystujące systemy regułowe;
- Metody oparte na reprezentacjach obiektowych, takich jak ramy, sieci semantyczne, ontologie;
- Metody używające modeli obliczeniowych.

Oprócz metod symbolicznych, wyróżnia się także reprezentacje niesymboliczne. Metody te odnoszą się do obserwacji i doświadczeń otaczającego świata. Przykładowo sztuczne sieci neuronowe, symulują cechy reprezentacji wiedzy i jej przetwarzania w komórkach nerwowych organizmów żywych. Wiedza zgromadzona jest w sposobie połączeń między neuronami oraz wartościach wag reprezentujących siłę tych połączeń. Do innych technik reprezentacji wiedzy należą tzw. algorytmy genetyczne, które umożliwiają przekazywanie wiedzy o gatunku następnym generacjom. Wiedza zapisana jest tutaj w tzw. chromosomach. Sposób reprezentacji danych powinien posiadać dwie podstawowe właściwości:

- Efektywność - pozwalającą na łatwą analizę danych w systemie komputerowym;
- Uniwersalność - umożliwiającą przechowywanie zbiorów danych różnego typu, opisujących badane procesy i zjawiska.

Spośród metod reprezentacji danych i wiedzy w rozprawie scharakteryzowano pokrótce tablicowy zapis danych, regułową reprezentację wiedzy oraz ontologie, ze względu na ich wykorzystanie w omawianych metodach.

2.3.1 Zbiory danych i ich reprezentowanie

Jedną ze struktur, które mogą być zastosowane do reprezentacji i przechowywania danych jest często wykorzystywany w praktyce tablicowy sposób reprezentacji danych. W podejściu tym dane przedstawiane są w postaci tablicy, w której każdy wiersz reprezentuje informacje na temat pojedynczego obiektu świata rzeczywistego i z tego powodu określany jest obiektem. Kolumny opisują cechy obiektu wyrażone za pomocą wartości numerycznych lub symbolicznych i określane są mianem atrybutów. Na przecięciu wierszy i kolumn znajdują się wartości poszczególnych atrybutów dla danych obiektów. Strukturę zdefiniowaną w ten sposób nazywa się systemem informacyjnym SI (ang. *information system*) lub rzadziej tablicą informacyjną lub tablicą typu atrybut-wartość [104].

Definicja 2.3.1 (System informacyjny SI) *System informacyjny to para postaci:*

$$\boldsymbol{SI} = (U, A), gdzie:$$

- *U jest niepustym, skończonym zbiorem zwanym uniwersum, przy czym elementy zbioru U nazywane są obiektami: $U = \{u_1, u_2, .., u_n\}$,*
- *A jest niepustym, skończonym zbiorem atrybutów: $A = \{a_1, a_2, .., a_m\}$,*

Zbiór V_a nazywa się dziedziną atrybutu $a \in A$, $V = \cup_{a \in A} V_a$. Definiuje się również funkcję informacyjną $f : U \times A \to V$ taką, że: $\forall_{u \in U, a \in A} f(u, a) \in V_a$.

Jeżeli jeden z atrybutów reprezentuje przynależność każdego obiektu do kategorii, wówczas mówimy o tablicy decyzyjnej DT (ang. *decision table*).

Definicja 2.3.2 (Tablica decyzyjna DT) *Tablicą decyzyjną nazywamy system informacyjny postaci:*

$$\boldsymbol{DT} = (U, A \cup \{d\}), gdzie:$$

- $d \notin A$ *jest atrybutem decyzyjnym nie należącym do zbioru atrybutów* A,
- *atrybuty* $a \in A$ *nazywa się atrybutami warunkowymi.*

Wartości atrybutu decyzyjnego dzielą zbiór obiektów na predefiniowane klasy, składające się z obiektów o tej samej wartości atrybutu decyzyjnego. Dla pojedynczych klas (odnoszących się do pewnych pojęć) definiuje się przykłady pozytywne (obiekty należące do tej klasy) i negatywne (przynależące do innych klas).

Celem klasyfikacji jest zbudowanie klasyfikatora, np. w formie drzewa decyzyjnego lub zbioru reguł klasyfikacyjnych, potrafiącego rozpoznawać przynależność nowych obiektów do odpowiedniej klasy w oparciu o wyuczone wcześniej wzorce. W wyniku klasyfikacji obiekt zostaje przyporządkowany do (wybranej przez klasyfikator) klasy. Klasyfikator służy więc do predykcji wartości atrybutu decyzyjnego obiektów, dla których wartość ta nie jest znana i może być traktowany jako przybliżony opis pojęć (klas decyzyjnych).

2.3.2 Regułowa reprezentacja wiedzy

Pośród metod reprezentacji wiedzy istotną rolę w praktycznych zastosowaniach odgrywają metody oparte na regułach. Jest to jedna z najstarszych metod reprezentacji wiedzy i jednocześnie najbardziej popularna [106, 72]. Ogólną postać takiej reprezentacji przedstawia wzór 2.1:

$$\text{JEŻELI przesłanka (warunek) TO wniosek (konkluzja)} \tag{2.1}$$

co oznacza, że jeśli przesłanka jest prawdziwa, to prawdziwa jest również konkluzja. Przesłanki definiują więc pewien wzorzec lub wymogi, których spełnienie pozwala na przyjęcie wniosku. Działanie reguły odbywa się według wywodzącej się z logiki reguły wnioskowania, tj. reguły odrywania (*modus ponens*) według której, jeżeli p implikuje logicznie q oraz p jest prawdziwe, to q jest również prawdziwe:

$$\frac{\begin{array}{l} p \rightarrow q \\ p \end{array}}{q}$$

gdzie p i q to litery zdaniowe. Jeżeli przesłanka reguły jest prawdziwa to mówi się, że reguła jest spełniona.

Warunki mogą przyjmować postać deskryptorów (selektorów postaci np. $a = v$) opartych na wybranych atrybutach [39], gdzie atrybut $a \in A$ oraz $v \in V_a$. W przypadku danych symbolicznych najczęściej stosuje się takie rodzaje deskryptorów jak: deskryptory równościowe ($a = v$) i podzbiorowe ($a \in \{v_1, ..., v_k\}$), natomiast w przypadku danych ciągłych: nierównościowe (np. $a < v$) i przedziałowe ($a \in [v_1, ..., v_k]$). Jeżeli część warunkowa reguły zawiera warunki zbudowane na wartościach atrybutów opisujących obiekty, a wniosek określa przynależność obiektu spełniającego te warunki do pewnego podzbioru nazywanego klasą (decyzyjną) lub pojęciem, to mówi się o regułach decyzyjnych.

Przesłanka może zawierać pewną liczbę deskryptorów połączonych funktorami logicznymi: koniunkcji *ORAZ (AND)* lub alternatywy *LUB (OR)*, jak w przykładzie 2.3.1.

Przykład 2.3.1 *Przykład reguły decyzyjnej.*

JEŻELI (A=x) ORAZ (B=y) TO (D=d)

gdzie A, B to atrybuty warunkowe, D jest atrybutem decyzyjnym, $x \in V_A$, $y \in V_B$, $d \in V_D$. Zapis oznacza, że dla pewnego obiektu i atrybutu A funkcja informacyjna przyjmuje wartość x i jednocześnie dla tego samego obiektu i atrybutu B funkcja informacyjna przyjmuje wartość y, to dla tego obiektu i atrybutu D funkcja przyjmuje wartość d.

Jeżeli warunki w złożonej przesłance są połączone funktorami koniunkcji, to proces analizowania takiej reguły jest kończony, z wynikiem negatywnym, po napotkaniu pierwszego niespełnionego warunku. Z tego powodu kolejność warunków w przesłance może mieć znaczenie dla prostoty obliczeń, chociaż wynik wnioskowania jest niezależny od tej kolejności.

Regułę zawierającą w części przesłankowej spójnik LUB można zastąpić zestawem równoważnych reguł bez tego spójnika, np. regułę 2.2:

$$\text{JEŻELI (A=x) LUB (B=y) TO (D=d)} \tag{2.2}$$

można zastąpić dwiema regułami:

$$\begin{gathered}\text{JEŻELI (A=x) TO (D=d)} \\ \text{JEŻELI (B=y) TO (D=d)}\end{gathered} \tag{2.3}$$

Dopuszczalna jest także tzw. *pełna* (rozwinięta) postać reguł zawierająca dodatkowe stwierdzenie uznawane za prawdziwe w przypadku niespełnienia przesłanki. Ogólną postać reguły pełnej przedstawia wzór 2.4.

$$\text{JEŻELI przesłanka TO konkluzja}_1 \text{ WPP konkluzja}_2 \tag{2.4}$$

przy czym konkluzja$_2$ jest dodatkowym stwierdzeniem, WPP oznacza "w przeciwnym przypadku". Ogólna postać reguł może jednak prowadzić do uznania nieoczekiwanych konkluzji.

Reguły mogą być charakteryzowane przez różne miary atrakcyjności, w tym stopień pewności CF (ang. *Certainty Factor*) czy współczynnik ufności CNF (ang. *Confidence Factor*) [134]. Zwykle są to liczby z przedziału [-1,1] lub [0,1] określające stopień przeświadczenia użytkownika co do pewności konkluzji, gdy przesłanka reguły jest spełniona.

Stosowany jest także bardziej formalny zapis reguł, gdzie opuszcza się symbol *JEŻELI*, a w miejsce słowa *TO* używa się symbolu implikacji. Przesłanki natomiast połączone są za pomocą funktorów logicznych pisanych w postaci symbolicznej. Reguła z przykładu 2.3.1 może być zapisana w postaci:

$$(A = x) \wedge (B = y) \Rightarrow (D = d) \tag{2.5}$$

Reguły można wykorzystywać do reprezentacji zależności pomiędzy pojęciami. Nie jest istotna dziedzina lecz charakter opisywanych powiązań. Ze względu na swoje zalety, takie jak naturalny sposób przedstawienia wiedzy i relatywnie niski koszt, regułowa reprezentacja wiedzy ma zastosowanie do reprezentacji wiedzy dziedzinowej.

2.3.3 Ontologie

Jednym ze sposobów reprezentacji wiedzy jest reprezentacja w postaci ontologii. Ontologia jest zwykle rozumiana jako skończony zbiór pojęć tworzących hierarchię i relacje między pojęciami z różnych poziomów hierarchii.

Słowo ontologia było pierwotnie używane w filozofii, gdzie oznacza analizę pojęć i idei celem ustalenia co istnieje oraz jakie są związki między istniejącymi elementami. Teoria ontologii wiąże się już z pracami Arystotelesa, G. Leibniza czy I. Kanta. Większość z nich traktuje ontologię jako naukę o rodzajach i strukturach obiektów, ich właściwości, zdarzeń, procesów czy relacji [87, 140, 62, 90]. W informatyce tego pojęcia używa się od lat 60 XX w. jako sposobu formalizacji wiedzy, głównie w kontekście rozwoju baz danych i sztucznej inteligencji.

W zastosowaniach informatycznych głównym celem tworzenia ontologii jest dzielenie się wiedzą w taki sposób, aby była zrozumiała i z łatwością przetwarzana przez człowieka jak i przez systemy informatyczne. Ontologie wykorzystują teorie wywodzące się z algebry, teorii zbiorów, sieci semantycznych oraz rachunków logicznych.

Definicja ontologii

W podejściu filozoficznym, jak również w zastosowaniach informatycznych, brak jest porozumienia, jeśli chodzi o definicję ontologii. Rozważmy trzy definicje ontologii, dobrze znane z literatury.

Wg Guarino [62] ontologia odnosi się do "specyficznego słownictwa używanego do opisania pewnej rzeczywistości (lub jakiejś części rzeczywistości), a także szeregu wyraźnych założeń dotyczących zamierzonego znaczenia słów ze słownika". W tym podejściu, ontologia opisuje hierarchię pojęć powiązanych relacjami, natomiast w bardziej skomplikowanych przypadkach, dodawane są odpowiednie aksjomaty do wyrażania innych relacji między pojęciami i ograniczania interpretacji tych pojęć.

Jedną z częściej przytaczanych definicji ontologii jest definicja sformułowana przez Grubera [61], będąca jedną z pierwszych definicji stworzonych na potrzeby informatyki. Definiuje on ontologię jako "formalną, jednoznaczną specyfikację dzielonej (wspólnej) konceptualizacji". W stwierdzeniu tym, konceptualizacja odnosi się do abstrakcyjnego modelu pewnego zjawiska lub bytu, który identyfikuje odpowiednie pojęcia rzeczywistego obiektu. Definicję tę zastosowano w dziedzinie sztucznej inteligencji w celu ułatwienia współdzielenia i ponownego użycia zgromadzonej wiedzy.

Kolejną jest definicja ontologii zalecana przez organizację World Wide Web Consortium (W3C) [161], według której "ontologia definiuje terminy używane do opisywania i przedstawiania obszaru wiedzy".

Centralnym pojęciem większości ontologii są klasy obiektów umożliwiające opis pojęcia w danej dziedzinie wiedzy. Pojęcie (ang. *concept*) oznacza ogólne określenie oznaczające zbiór obiektów posiadających wspólne właściwości, którymi odróżniają się od innych pojęć. Pojęcia służą człowiekowi do myślenia o rzeczywistości, pozwalają na zmniejszenie ilości przetwarzanych informacji w jak najkrótszym czasie, np. poprzez przydzielenie (zaklasyfikowanie) danego obiektu do znanej wcześniej klasy. Klasyfikacja to naturalny sposób rozpoznawania rodzaju rzeczy lub zjawisk. Pojęcia opisywane są jako podstawowa struktura poznawcza reprezentująca uogólnioną klasę obiektów (przedmiotów, zdarzeń, czynności, relacji) (patrz [38]) i stanowią jeden z rodzajów reprezentacji (przedstawienia). Rola pojęć w procesie poznania świata jest ogromna. Pojęcia są narzędziami, za pomocą których człowiek poznaje świat i ujmuje zdobytą wiedzę. Pojęcia pełnią więc rolę repozytoriów wiedzy. Ludzie posługują się hierarchiami pojęć. Formalnymi cechami pojęć są ogólność i abstrakcyjność. Ogólność oznacza powiązanie cech, za pomocą których rozum odnosi się do wielu obiektów jednostkowych, natomiast abstrakcyjność polega na pomijaniu większości cech przysługujących jednostkom.

Struktura ontologii

Ontologie pojęć łączy wiele podobieństw strukturalnych, niezależnie od języka, w którym są wyrażone. Większość ontologii opisuje obiekty (instancje), pojęcia (klasy), atrybuty (właściwości) i relacje (patrz np. [61, 62, 73, 161]).

Obiekty (instancje) są podstawowymi komponentami bazowego poziomu ontologii. Mogą obejmować rzeczywiste obiekty, takie jak ludzie, zwierzęta, samochody, rośliny czy planety, a także obiekty abstrakcyjne, jak numery i słowa.

Pojęcia (klasy) są abstrakcyjnymi grupami lub zbiorami obiektów. Mogą zawierać obiekty lub inne pojęcia. Przykładami pojęć są: pojazd (klasa wszystkich urządzeń technicznych służących o przemieszczania się ludzi), pacjent (klasa wszystkich osób leczonych), nadciśnienie (klasa wszystkich pacjentów cierpiących z powodu nadciśnienia) czy zespół (klasa wszystkich graczy z jakiegoś zespołu).

Instancje należące do pojęć w ontologii można opisać poprzez zestawy wartości wybranych cech (atrybutów). Każdy atrybut posiada co najmniej nazwę oraz wartość, i jest wykorzystywany do przechowywania informacji charakterystycznych dla obiektu, dla którego atrybut jest przeznaczony. Na przykład, obiekt pojęcia *Uczestnik* posiada atrybuty, takie jak imię, nazwisko, adres zamieszkania, przynależność. Jeśli nie zdefiniuje się atrybutów pojęć, wówczas mówi się o taksonomii (gdy opisane są relacje pomiędzy pojęciami) lub kontrolowanym słowniku. Są one użyteczne, lecz nie są uważane za prawdziwe ontologie.

Wyróżnia się trzy typy relacji między pojęciami z ontologii:

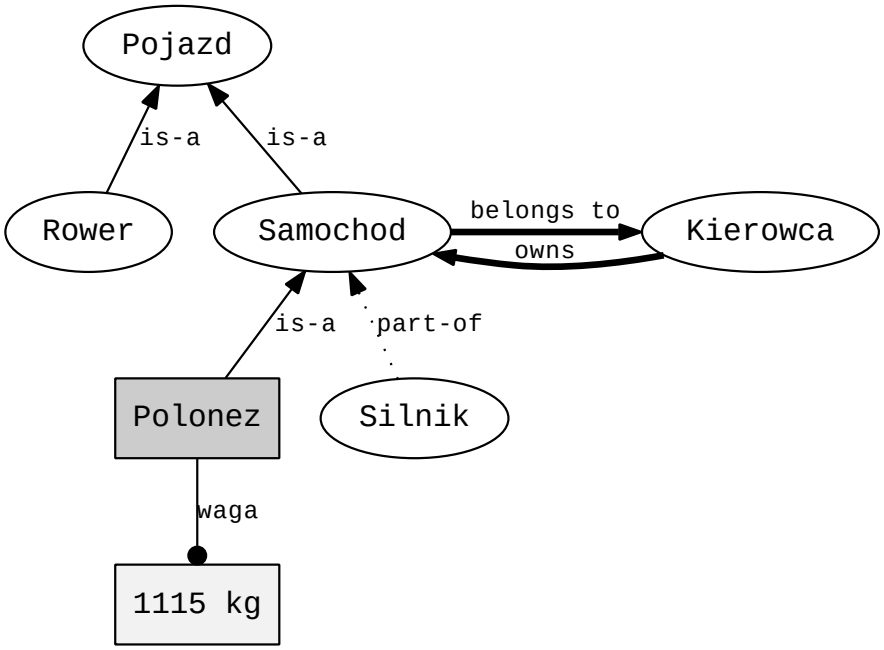
- Relacja subsumcji - oznaczana jako relacja `'jest'` (ang. `'is-a'`);
- Relacja meronimii - inaczej relacja część-całość, oznaczana jako `'jest częścią'` (ang. `'part-of'`);
- Relacja specyficzna dla danej dziedziny.

Pierwszym typem relacji jest *relacja subsumcji*, inaczej przynależności. Jeżeli pojęcie B jest w relacji subsumcji, tzn. jest podporządkowane pojęciu A, to mówi się, że B jest rodzajem A, co jest tożsame ze zwrotem: *klasa B jest podklasą A*. Jeżeli klasa B jest podklasą A, to klasę A nazywa się nadklasą. Relacja subsumcji jest bardzo podobna do pojęcia dziedziczenia, dobrze znanego z programowania obiektowego. Taki związek można stosować do tworzenia hierarchii pojęć, zwykle zawierającej najbardziej uogólnione pojęcia takie jak *Pojazd* na górze hierarchii, a bardziej szczegółowe, jak *Samochód* na dole. Hierarchia pojęć zwykle przedstawiana jest za pomocą grafu ontologii (przykład na Rys. 2.1), w którym relacja subsumcji reprezentowana jest za pomocą cienkiej, ciągłej linii ze strzałką skierowaną w kierunku od podklasy do nadklasy.

Innym powszechnym typem relacji jest *relacja meronimii*, która prezentuje jak obiekty łączą się razem, tworząc obiekty złożone. Mianem meronimu nazywa się

część składową lub fragment większej całości, czyli holonimu. W zwrocie *A jest częścią B*, pojęcie *A* to meronim, *B* to holonim. Na przykład, w ontologii z Rys. 2.1, *Silnik* jest częścią *Samochodu*. Relacja meronimii jest przedstawiana graficznie linią przerywaną ze strzałką w kierunku od części do obiektu złożonego.

Oprócz standardowych relacji `'jest'` i `'jest częścią'`, ontologie często zawierają dodatkowe typy relacji, które dalej udoskonalają semantykę modelowaną przez ontologię. Relacje te są często specyficzne dla dziedziny i są wykorzystywane do odpowiedzi na szczegółowe rodzaje pytań. Na przykład, w dziedzinie pojazdów, można zdefiniować relację `'należy do'` (ang. `'belongs to'`) pomiędzy pojęciami *Samochód* i *Kierowca*, która określa kierowcę będącego właścicielem samochodu. W dziedzinie pojazdów, definiuje się również relację `'posiada'` (ang. `'owns'`) między pojęciami *Kierowca* i *Samochód*, która mówi, który kierowca posiada dany samochód. Relacje tego typu są reprezentowane przez grubą, ciągłą linię ze strzałką. Przykładową, prostą ontologię przedstawia Rys. 2.1, w której zaprezentowano wszystkie trzy typy relacji między pojęciami. Linia ciągła z kropką reprezentuje atrybut obiektu *Polonez* o nazwie *waga* i wartości 1115 kg.



Rysunek 2.1: Przykładowa ontologia.

Budowa ontologii - ogólne zalecenia

Istnieje szereg doniesień opisujących doświadczenia różnych grup projektantów, uzyskane w procesie budowy ontologii (patrz np. [75]). Chociaż nie dostarczają one

jeszcze formalnych ram umożliwiających tworzenie zintegrowanej metodologii, na ich podstawie mogą być tworzone ogólne zalecenia dotyczące tworzenia ontologii. Każdy projekt związany z tworzeniem ontologii składa się z następujących etapów:

- Określenie motywacji do tworzenia ontologii - istotna dla całego procesu jest jasność celu, dla którego ontologia ma być budowana;
- Ustalenie domeny oraz zasięgu ontologii, czyli określenie jakiego wycinka modelowanego świata będzie dotyczyła, np. za pomocą tzw. pytań kompetencyjnych [160]. Punktem wyjścia dla tej metody jest określenie listy pytań, na które powinna odpowiadać baza danych utworzona na podstawie ontologii;
- Tworzenie ontologii:
  - Tworzenie słownika zawierającego terminy używane przez ontologię, jak również ich definicje;
  - Identyfikacja pojęć (klas);
  - Tworzenie struktury (hierarchii) pojęć;
  - Modelowanie relacji między pojęciami ontologii;
- Ocena otrzymanej ontologii;
- Implementacja ontologii.

Wśród podejść do budowy hierarchii klas wykorzystuje się takie podejścia, jak trzy podane w artykule [159]:

1. Góra-dół (ang. *top-down*) - zaczyna się od pojęcia nadrzędnego do wszystkich pojęć zawartych w bazie wiedzy i przechodzi się do następnych poziomów niższych pojęć przez zastosowanie atomizacji;
2. Dół-góra (ang. *bottom-up*) - zaczyna się od pojęć najniższego poziomu i przechodzi do pojęć na wyższych poziomach hierarchii stosując uogólnienia;
3. Od środka (ang. *middle-out*) - zaczyna się od pojęć, które są najbardziej istotne z punktu widzenia projektu i w zależności od potrzeby stosuje się atomizację lub uogólnienia.

Ontologie tworzone na potrzeby aplikacji komputerowych wymagają formalnego języka, przy pomocy którego można je budować i przechowywać. Do standardów zapisu ontologii należą technologie oparte na bazie języka XML (ang. *eXtensible Markup Language*), takie jak: Resource Description Framework (RDF) utworzony przez konsorcjum World Wide Web Consortium (W3C), F-logic czy Web Ontology Language (OWL) oparty na DAML+OlL (DARPA Agent Markup Language+Ontology Inference Layer) [144, 90]. Do zapisania i przechowywania ontologii utworzonej na potrzeby rozprawy wykorzystano technologię OWL.

2.4 Definicja wiedzy dziedzinowej

Celem podjęcia próby zdefiniowania wiedzy dziedzinowej, należy przybliżyć czym jest wiedza. W literaturze wskazuje się, że podstawą wiedzy są dane i informacja, które wiedzą stają się dopiero po ich przetworzeniu [29]. Dane to surowe fakty, niezinterpretowane wartości sygnałów, które otrzymujemy np. z urządzeń pomiarowych. Informacjami są dane z przypisanym znaczeniem. Natomiast wiedza w dużym uproszczeniu, oznacza ogół danych i informacji, które ludzie wykorzystują w praktyce do wykonywania działań i tworzenia nowych informacji. Charakterystyka przedstawionych pojęć w literaturze specjalistycznej często uzupełniana jest tezą, że tworzą one pewien łańcuch pojęć uzupełniany mądrością: dane - informacja – wiedza – mądrość. Prosty ilustrację wymienionych pojęć stanowi Przykład 2.4.1.

Przykład 2.4.1 *Przykład danej, informacji i wiedzy.*

***Dane**: liczba 31*
***Informacja**: liczba 31 to wartość wskaźnika masy ciała BMI (ang. body mass index) wyznaczanego ze wzoru: $BMI = \frac{waga}{wzrost^2}[\frac{kg}{m^2}]$*
***Wiedza**: liczba 31 to wartość wskaźnika BMI oznaczająca otyłość*

Wiedza przedstawiona w Przykładzie 2.4.1 może być wykorzystana do podjęcia działań mających na celu obniżenie masy ciała. Umiejętność skorzystania z posiadanej wiedzy bywa nazywana mądrością [147].

Z tych definicji danych, informacji i wiedzy, można stwierdzić, że różnice między nimi nie są ściśle określone ani statyczne. Wynika to z faktu, że wiedza w dużej mierze zależy od kontekstu. Często trudno jest odróżnić wiedzę od informacji, gdyż dla jednych informacja pozostaje informacją, a u innych przekształca się w wiedzę. Wiedza informatyka nie ma większego sensu dla np. biologa, ponieważ biolog nie zna się dobrze na informatyce. W tym sensie wiedza informatyka stanowi dane dla biologa. Pojęcia te są więc wymienne, a dane lub informacje stają się wiedzą, kiedy przypisywane jest im znaczenie oraz cel.

Termin 'wiedza' występuje zarówno w języku potocznym, jak i na gruncie wielu dyscyplin naukowych, takich jak filozofia, psychologia czy informatyka i w każdym przypadku przypisywane mu jest nieco inne znaczenie. Wiedza jest pojęciem, dla którego nie zdefiniowano dotychczas jednej, satysfakcjonującej i akceptowanej przez różnych specjalistów definicji.

Ogólnie przez wiedzę rozumie się ogół utrwalonych wiadomości i umiejętności z jakiejś dziedziny (doświadczenie) wraz ze zdolnością ich interpretacji, czyli analizy i wnioskowania w celu praktycznego wykorzystywania [81]. Wiedza jest pojęciem bardzo szerokim, dlatego istnieje wiele jej podziałów i klasyfikacji. Wśród kry-

teriów podziału wiedzy wyróżnia się przedmiot, którego dotyczy (np. wiedza społeczna, medyczna), jej pochodzenie (np. wiedza empiryczna, aprioryczna), zasięg (np. wiedza specjalistyczna), status poznawczy (np. wiedza naukowa, potoczna, teoretyczna), okres występowania czy lokalizację geograficzną.

Istotnym podziałem w kontekście pozyskiwania i wykorzystania wiedzy jest podział na wiedzę jawną i ukrytą. Wiedza jawna (ang. *explicit knowledge*), zwana formalną, obiektywną lub uzewnętrznioną, jest wiedzą usystematyzowaną i wyraźnie sprecyzowaną. Wyrażana jest w formie słów, liczb i symboli, jako dokumenty (instrukcje, procedury, transakcje, raporty, regulaminy) lub dane w systemach informatycznych.

Wiedza ukryta (ang. *tacit knowledge*), nazywana cichą, jest trudna do jasnego sprecyzowania. Gromadzi się wraz ze wzrostem doświadczenia i przekazywana jest głównie w formie werbalnej. Wiedza ta nazywana jest ukrytą, ponieważ przechowywana jest w umysłach i świadomości, czyli pamięci ludzi, którzy ją wytworzyli lub pozyskali. Występuje w wielu trudnych do określenia postaciach i ujawnia się w efektywnym sposobie rozwiązywania problemów przez ekspertów. Trudności w jej wyrażeniu wynikają z ogromnej liczby wyróżnialnych stanów w obserwowanej rzeczywistości i ograniczoności języka, stosowanego przy jej abstrakcyjnym odwzorowaniu. W większości organizacji zasoby wiedzy ukrytej są znacznie większe od zasobów wiedzy jawnej. Szacuje się, iż około 80% wiedzy zgromadzonej w przedsiębiorstwach ma charakter wiedzy ukrytej [42].

Istnieje także drugi rodzaj wiedzy ukrytej. Jest to potencjalna wiedza zawarta w gromadzonych różnego typu dokumentach i bazach danych. Z faktów i informacji w nich zawartych możliwe jest pozyskanie wiedzy. Identyfikacja tej wiedzy ukrytej i wytwarzanie na jej bazie zasobów wiedzy jawnej jest głównym celem metod eksploracji danych, która umożliwia odkrycie zupełnie nowej wiedzy nie znanej wcześniej nawet specjalistom i ekspertom dziedzinowym. Należy podkreślić, że cała wiedza jawna ma swoje źródło w wiedzy ukrytej. Wiedza najpierw powstaje w umysłach ludzi jako ukryta, a dopiero później jest formalizowana i wyrażana za pomocą metod reprezentacji wiedzy, stając się wiedzą jawną.

Wiedza dziedzinowa jest jednym z rodzajów wiedzy, wyróżnionej na podstawie zasięgu jej występowania do pewnej dziedziny. Dla kontrastu, wiedza, która funkcjonuje skutecznie w każdej dziedzinie nazywana jest wiedzą niezależną od dziedziny (ang. *domain-independent knowledge*). Pod pojęciem wiedzy dziedzinowej kryje się zatem wiedza, jaką posiadają specjaliści w różnych dziedzinach, tj. lekarze czy ekonomiści. Opiera się ona na wielu skojarzeniach pomiędzy przyczynami obserwowanych danych i faktów. Ekspert wyposażony w taką wiedzę, nazywany ekspertem dziedzinowym często używa metod heurystycznych do rozwiązywania problemów probabilistycznych oraz wykorzystuje w procesie decyzyjnym błędne dane. Rozwija on swoją wiedzę przez lata doświadczeń przy rozwiązywaniu proble-

mów w wąskiej dziedzinie, uczy się, wykorzystując zdobyte doświadczenie, modyfikuje zbiór swoich pojęć, kieruje się zdrowym rozsądkiem, ma intuicję i rozumuje na podstawie analogii. Ekspert zatem to człowiek posiadający specjalistyczną wiedzę z określonej dziedziny i umiejętność stosowania jej do rozwiązywania problemów z tej dziedziny [72].

Zastosowanie wiedzy dziedzinowej w procesie KDD ma odzwierciedlać proces uczenia się człowieka. Od dzieciństwa człowiek nabywa wiedzę albo metodą prób i błędów, lub poprzez edukację. W obliczu nowych zadań, jest on w stanie efektywnie wykorzystać zdobytą wiedzę do poprawienia swoich umiejętności. Podczas uczenia się pojęć, człowiek wykorzystuje nie tylko dostępne przykłady uczonego pojęcia, ale także wcześniejszą wiedzę [167, 67]. Brak informacji w jednym źródle jest kompensowany przez inne źródło. Wiedza posiadana przez człowieka wpływa na jego interpretację przykładów. Dobór przykładów uczonego pojęcia, jest u ludzi oparty na wcześniejszej wiedzy i odwrotnie, przykłady wpływają na wiedzę. Tak więc jeżeli proces KDD ma naśladować zdolność człowieka do nabywania wiedzy, musi posiadać możliwość zastosowania zdobytej uprzednio wiedzy do procesu odkrywania wiedzy. Jeśli wiedza dziedzinowa już istnieje, proces KDD nie powinien jej ignorować i zaczynać poszukiwań od stanu zerowego. Nie powinien też na nowo odkrywać istniejącej już wiedzy dziedzinowej. W szczególności, przy rozwiązywaniu rzeczywistych problemów, wcześniejsza wiedza jest na tyle cenna, że należy ją włączyć do praktycznych systemów KDD. W literaturze opisywany jest nowy paradygmat eksploracji danych oparty m.in. na wiedzy dziedzinowej (ang. *domain-driven*), w przeciwieństwie do dotychczasowego opartego tylko na danych (ang. *data-driven*) [36].

Wiedza dziedzinowa zastosowana w procesie KDD ma na celu naprowadzanie poszukiwań na interesujące obszary. Umożliwia w ten sposób zmniejszenie przestrzeni poszukiwań czy redukcję liczby odkrywanych wzorców. Ułatwia także identyfikację i interpretację otrzymanych w procesie KDD wyników, co ma zasadnicze znaczenie dla zamieniania wzorców w ciekawą, zrozumiałą i praktyczną wiedzę [93]. Gdy generowany jest zbiór wzorców, wiedza dziedzinowa może pomóc użytkownikowi określić, jak dobrze wzorce te pasują do istniejącej wiedzy, czy są z nią zgodne lub jej zaprzeczają [97]. Dzięki tym wszystkim czynnikom, możliwe jest uzyskanie lepszej wydajności procesu KDD. Należy zdawać sobie jednak sprawę z braku możliwości zastosowania całej dostępnej wiedzy dziedzinowej dla jednego zadania. Wiedza ta bowiem jest kontekstowo zależna.

W literaturze można spotkać wiele definicji wiedzy dziedzinowej wykorzystywanej w KDD. Jedna z nich określa wiedzę dziedzinową jako wszystkie dostępne informacje dotyczące zadania do wyuczenia dodane do przykładów treningowych [131]. Według [8] są to informacje na temat danych pochodzące z innego procesu odkrywania wiedzy lub od ekspertów dziedzinowych. W [2] przedstawiono ją

w postaci pewnego rodzaju porad, pobocznych informacji, heurystyk czy formalnych reguł. Wiedza dziedzinowa jest tam definiowana jako dodatkowa informacja na temat funkcji celu stosowana do kierowania procesem uczenia się. Wiedza dziedzinowa definiowana jest także jako wszelkie informacje, które nie są jawnie zaprezentowane w systemie.

Przykładami wiedzy dziedzinowej w zagadnieniach medycznych może być następująca wiedza: *Pacjenci płci męskiej nie mogą mieć rozpoznania ciąży*, *Pacjenci płci żeńskiej nie mogą mieć rozpoznania nowotworów prostaty*. Inną formą tej wiedzy jest uogólnianie wartości atrybutów na jej podstawie, np. atrybut *Wiek* może zostać podzielony na przedziały wiekowe takie jak: {młody, w średnim wieku, stary}. Uogólnianie dziedziny atrybutów daje w wyniku bardziej zrozumiałe, a więc bardziej użyteczne wzorce. Reguła postaci: *Jeżeli pacjent urodził się w dniu: 01.05.2001, to nie ma nadciśnienia* jest mniej użyteczna dla lekarzy niż reguła: *Jeżeli pacjent jest młody, to nie ma nadciśnienia*. Na gruncie zastosowań biznesowych takimi przykładami są: *Klienci z dużym przychodem mają niskie ryzyko kredytowe*.

Reprezentacja wiedzy dziedzinowej za pomocą reguł. Wiedza dziedzinowa może być reprezentowana w różny sposób. Często zakodowana jest w postaci reguł logicznych. Formalnie, wiedza dziedzinowa (WD) może być reprezentowana w postaci reguły 2.6:

$$WD = \{X \Rightarrow Y\} \tag{2.6}$$

gdzie X i Y stanowią proste lub połączone koniunkcją warunki dotyczące atrybutów tablicy decyzyjnej. Załóżmy, że chcemy się dowiedzieć, czy pewien lek X ma wpływ na pacjentów chorujących na grypę. I załóżmy, że dostępna wiedza dziedzinowa (WD) obejmuje następujące stwierdzenia: *Personel medyczny pewnego szpitala S został poddany szczepieniu przeciw grypie* oraz *Osoby poddane szczepieniu przeciw grypie nie chorują na grypę*, zapisane w postaci:

$$\begin{aligned}&\text{(Miejsce zatrudnienia=szpital } S) \Rightarrow \text{(Szczepienie przeciw grypie=TAK)}\\ &\text{(Szczepienie przeciw grypie=TAK)} \Rightarrow \text{(Zachorowanie na grypę=NIE)}\end{aligned} \tag{2.7}$$

Na podstawie dostępnej wiedzy WD możliwe jest wyprowadzenie pochodnej wiedzy dziedzinowej (PWD). Na przykład, poprzez wykorzystanie przechodniej zależności, można ustalić nową wiedzę dziedzinową w postaci stwierdzenia: "personel medyczny szpitala S nie zachoruje na grypę", reprezentowanego przez regułę:

$$\text{(Miejsce zatrudnienia=szpital } S) \Rightarrow \text{(Zachorowanie na grypę=NIE)} \tag{2.8}$$

Niech WD będzie zbiorem całej wiedzy dziedzinowej dostępnej dla danego problemu. Definiuje się WD^+, domknięcie WD [103], jako:

$$WD^+ = WD \cup \{PWD_i | PWD_i \text{ jest wyprowadzalna z WD}\} \tag{2.9}$$

Oznacza to, że zbiór całej wiedzy dziedzinowej składa się z tej określonej przez eksperta dziedzinowego oraz tej, która została wyprowadzona ze zdefiniowanej wiedzy dziedzinowej.

Reprezentacja wiedzy dziedzinowej za pomocą ontologii Jednym ze sposobów przedstawiania wiedzy dziedzinowej jest zapis w postaci ontologii pojęć, gdzie ontologia jest zwykle rozumiana jako skończona hierarchia pojęć i relacji łączących pojęcia z różnych poziomów (patrz [62]). Obecnie, ontologie są stosowane jako alternatywny model reprezentacji wiedzy w wielu obszarach eksploracji danych, umożliwiając różne poziomy uogólniania pojęć oraz odkrywanie wzorców na różnych poziomach abstrakcji.

2.5 Rola wiedzy dziedzinowej w procesie odkrywania wiedzy

Wiedza dziedzinowa odgrywa kluczową rolę przede wszystkim w początkowych i końcowych etapach procesu odkrywania wiedzy z danych. Jednak doniesienia wskazują na jej pewną, chociaż zróżnicowaną rolę we wszystkich fazach projektu KDD [82, 52]. W rozdziale omówione zostaną kolejne etapy procesu KDD według podejścia w [92], ze wskazaniem możliwości zastosowania wiedzy dziedzinowej w każdym z nich.

(1) Zrozumienie dziedziny badań i określenie celów procesu powinno uwzględniać różne aspekty badanej dziedziny. Do realizacji tego celu wskazane jest uwzględnienie poglądów na temat problemu osób (ekspertów) zajmujących się daną dziedziną pod kątem różnych aspektów, tj. praktyków i teoretyków czy kadry zarządczej jak i sprzedawców.

(2) Wybór i utworzenie zbioru danych - musi uwzględniać cel procesu. Często jeden zbiór danych nie pokrywa wszystkich aspektów problemu badawczego. Adekwatne oraz uzupełniające się źródła danych mogą być wskazane przez eksperta dziedzinowego, takiego jak specjalista baz danych czy projektant systemu. Rola wiedzy dziedzinowej w tym etapie dotyczy także określenia struktury dostępnych informacji i ich wartości semantycznej ze wskazaniem ograniczeń występujących w danych.

(3) Wstępna obróbka i oczyszczanie danych - polega m.in. na eliminacji nieistotnych atrybutów i określeniu sposobu obchodzenia się z wartościami brakującymi (ang. *missing values*). Jedną z metod uzupełniania wartości brakujących danego atrybutu jest zastosowanie nadzorowanego algorytmu DM, w którym celem (decyzją) jest tenże atrybut. Wartości brakujące mogą być także wnioskowane na podstawie wartości powiązanych atrybutów wskazanych przez eksperta.

(4) Transformacja danych obejmująca metody redukcji ich wymiarów, takie jak selekcja czy ekstrakcja cech oraz transformację atrybutów, taką jak dyskretyzacja wartości atrybutów również wykorzystuje wiedzę dziedzinową. Na przykład, pewna informacja, która nie została zawarta w zbiorze danych, może zostać wywnioskowana na podstawie wartości innego atrybutu lub atrybutów przy ustaleniu pewnych założeń eksperta. Na tym etapie odbywać się może określenie skali czasu dla obserwacji rozłożonych w czasie, próbkowanie lub eliminacja przykładów (obiektów). Zastosowana tutaj wiedza dziedzinowa prowadzi do zmniejszenia przestrzeni poszukiwań i utworzenia zbioru danych, którego eksploracja daje trafniejsze wzorce.

(5) Wybór zadania eksploracji danych ze względu na cel eksploracji i typy odkrywanych wzorców, spośród takich klas jak: klasyfikacja i predykcja, grupowanie, inaczej analiza skupień (ang. *clustering*), odkrywanie asocjacji, analiza przebiegów czasowych, odkrywanie wzorców sekwencji, odkrywanie charakterystyk, opisy pojęć czy eksploracja tekstu.

Ważny jest odpowiedni dobór metody eksploracji do analizowanego zbioru informacji i oczekiwanych efektów. Wiedza dziedzinowa ma wpływ na ten etap poprzez określenie celu odkrywania wiedzy.

(6) Dobranie algorytmu DM - podobnie jak w poprzednim etapie nie istnieją konkretne wytyczne dotyczące wyboru algorytmu. Badacze przy wyborze kierują się takimi czynnikami jak: cel zadania, struktura zbioru danych, poprawność wyników, czas obliczeń, miary oceny jakości wyników czy dobór parametrów. Nie bez znaczenia jest także doświadczenie osoby wykonującej eksplorację danych w tym zakresie. Wybór algorytmu bywa tak skomplikowany, że wykorzystuje się więcej niż jedną technikę w celu osiągnięcia lepszych wyników.

(7) Eksploracja danych (DM) - budowa modelu na podstawie zebranych przypadków jest uważana za najważniejszą fazę w procesie KDD. W zależności od wybranego algorytmu, odkryta wiedza może być interpretowana przez ekspertów dziedzinowych, np. reguły zdefiniowane przez drzewo decyzyjne są sprawdzane przez eksperta. Na tym etapie algorytmy wykonywane są metodą prób i błędów, z różnym udziałem klas decyzyjnych w zbiorze treningowym, cech wejściowych czy z różnymi parametrami. Wydajność otrzymanych modeli wskazuje, które z modeli są najbardziej odpowiednie dla problemu decyzyjnego.

(8) Ewaluacja i interpretacja pozyskanych wzorców - ocena odkrytej wiedzy zazwyczaj polega na badaniu wydajności modelu przy użyciu danych testowych. Interpretacja wiedzy może być oparta na wydajności modelu oraz na sprawdzeniu, przejrzeniu odkrytej wiedzy. Kryteria oceny mogą być związane z celami biznesowymi i określone przez ekspertów w tej dziedzinie. Niepełna reprezentatywność danych treningowych może skutkować otrzymaniem modelu częściowo nieadekwatnego do problemu. Odczytanie takiego modelu przez eksperta i jego dostrojenie

na podstawie wiedzy dziedzinowej będzie skutkować lepszą jakością modelu w porównaniu z sytuacją gdy wiedza dziedzinowa nie była wykorzystana.

(9) Zastosowanie odkrytej wiedzy - podczas tej fazy odkryta wiedza jest łączona z dotychczasową wiedzą dziedzinową, by stać się częścią zbioru całej wiedzy dziedzinowej (WD^+). Wiedza dziedzinowa odgrywa ważną rolę podczas tego etapu. Eksperci mogą sugerować zbadanie obiektów wybranej klasy w sposób bardziej szczegółowy, np. w dodatkowych kategoriach i pod kątem atrybutów, które nie uczestniczyły w algorytmie klasyfikacji. Etap ten jest niestety często pomijany w projektach eksploracji danych jako wychodzący poza zakres eksperymentu DM. Ważnym aspektem jest także przekonanie użytkowników do stosowania otrzymanej wiedzy. Można to osiągnąć przez wyjaśnienie proponowanych wzorców lub wizualizację danych i odkrytej wiedzy.

Wiedza dziedzinowa ma zatem znaczenie na każdym etapie procesu KDD. Podsumowanie jej roli w KDD przedstawia Tab. 2.1.

Etap KDD	**Istotność WD**	**Rodzaj wiedzy dziedzinowej**
(1) Zrozumienie dziedziny badań i określenie celów procesu	Duża	wiedza jawna i ukryta
(2) Wybór i utworzenie zbioru danych	Średnia	relacje między atrybutami, semantyka baz danych
(3) Wstępna obróbka i oczyszczanie danych	Duża	wiedza jawna i ukryta
(4) Transformacja danych	Duża	wiedza jawna i ukryta, interpretacja wybranych cech
(5) Wybór zadania eksploracji	Średnia	
(6) Dobranie algorytmu DM	Średnia	
(7) Eksploracja danych, DM	Niska	sprawdzenie odkrytych wzorców
(8) Ewaluacja i interpretacja wzorców	Średnia	definicja kryteriów oceny
(9) Zastosowanie odkrytej wiedzy	Duża	dodatkowa wiedza dziedzinowa potrzebna do implementacji

Tablica 2.1: Rola wiedzy dziedzinowej w poszczególnych etapach procesu KDD.

Trzeba zwrócić uwagę, że w poszczególnych etapach procesu KDD zaangażowani są eksperci z wielu dyscyplin, które wydają się odgrywać kluczową rolę w efektywnym odkrywaniu wiedzy, np. eksperci baz danych, analitycy danych czy eksperci dziedzinowi. W rozprawie skoncentrowano się na zastosowaniu wiedzy ekspertów dziedzinowych.

2.5.1 Obszary zastosowań wiedzy dziedzinowej w odkrywaniu wiedzy

Wiedza dziedzinowa wspiera odkrywanie wiedzy poprzez koncentrowanie uwagi na wybranych aspektach problemu decyzyjnego. Celem jej zastosowania jest zmniejszenie przestrzeni poszukiwań poprzez np. redukcję rozmiaru zbioru danych, optymalizację hipotez reprezentujących wiedzę do odkrycia, weryfikację potencjalnie sprzecznych reguł czy zapobieganie tworzeniu reguł nadmiarowych [103]. W rozdziale omówione zostaną przykłady wykorzystania wiedzy dziedzinowej w wymienionych obszarach.

Redukcja przestrzeni poszukiwań. Rozmiar zbioru danych może zostać zredukowany przez wyeliminowanie nieistotnych atrybutów jak i obiektów, które nie są konieczne w odkrywaniu wiedzy. Rozważmy medyczny zbiór danych, w którym prosta wiedza dziedzinowa jest stwierdzeniem, że *Mężczyźni nie mogą być w ciąży.* Jeżeli chcemy odkryć, *Czy pewien lek X ma wpływ na ciążę*, wówczas WD może pomóc w zredukowaniu zbioru danych poprzez wyeliminowanie z rozważań przykładów pacjentów płci męskiej. Kolejna wiedza dziedzinowa, tj. *Kobiety poniżej 12 roku życia lub powyżej 65 roku życia nie zachodzą w ciążę* może być zastosowana do dalszej redukcji rozmiaru zbioru. Zatem wiedza dziedzinowa (WD) dla tego przykładu jest reprezentowana jako:

$$\begin{aligned} \text{WD} = \{&(\text{płeć} = \text{żeńska}) \Rightarrow (\text{ciąża} = \text{TAK}), \\ &(\text{wiek} > 12) \Rightarrow (\text{ciąża} = \text{TAK}), \\ &(\text{wiek} \leq 65) \Rightarrow (\text{ciąża} = \text{TAK}), ...\} \end{aligned} \tag{2.10}$$

Podstawową formą reprezentacji hipotezy (H) jest reprezentacja regułowa:

$$\text{H: JEŻELI przesłanka TO wniosek,} \tag{2.11}$$

gdzie przesłanka to warunek lub zestaw warunków czy kryteriów, sformułowanych przez ekspertów w danej dziedzinie celem zawężenia poszukiwań, a wniosek będzie stanowił odkrytą wiedzę, kiedy przesłanki zostaną spełnione (prawdziwe w zbiorze danych). Początkowa hipoteza w przykładzie może być więc przedstawiona następująco:

$$\text{H: JEŻELI (ciąża = TAK) ORAZ (terapia = X) TO (efekt = TAK)} \tag{2.12}$$

Oczywiście rzeczywista hipoteza do odkrycia może zawierać także inne atrybuty dotyczące pacjenta, jak np. waga, rasa, itd. Algorytm redukcji danych może zastosować WD do początkowej hipotezy w celu utworzenia zbioru ograniczeń. Mianowicie, dla każdego warunku w hipotezie, algorytm redukcji przeszukuje zbiór

wiedzy dziedzinowej. Jeżeli warunek znajduje się w części Y wiedzy dziedzinowej (patrz wzór 2.6 w Rozdz. 2.4), wówczas część X wiedzy WD jest wybierana jako ograniczenie. Zbiór takich ograniczeń wskazuje, które obiekty będą brane pod uwagę w procesie odkrywania wiedzy dla założonej hipotezy. W przykładzie, do analizowanego zbioru danych wejdą więc obiekty płci żeńskiej, w wieku powyżej 12 lat lub poniżej 65 lat.

Optymalizacja hipotezy. Wiedza dziedzinowa, poza zmniejszeniem rozmiaru danych może być również wykorzystana do określenia optymalnej hipotezy poprzez eliminację niepotrzebnych warunków w hipotezie. Taka optymalizacja skraca czas wyszukiwania interesującej wiedzy w danych. Na ogół w danych występują pewne zależności pomiędzy atrybutami jak i wewnątrz nich, co oznacza, że niektóre warunki mogą być implikowane przez inne. Zależności te można zidentyfikować m.in. za pomocą WD. W następstwie, warunki implikowane przez inne mogą być usunięte z hipotezy, ponieważ nie dostarczają żadnych dodatkowych informacji w odkrywaniu wiedzy, co powoduje przyspieszenie procesu odkrywania. Dla przykładu rozważmy problem rozpoznawania czynników, które wpływają na duże zużycie paliwa w samochodach. Dane opisane w [179] zawierają informacje na temat całkowitej długości samochodu (SIZE), liczby cylindrów (CYL), obecności turbosprężarki (TURBO), rodzaju układu paliwowego (FUELSYS), objętości skokowej silnika (DISPLACE), stopnia sprężania (COMP), mocy (POWER), rodzaju skrzyni biegów (TRANS), wagi (WEIGHT) oraz przebiegu auta (MILEAGE). Odkrywanie można rozpocząć od hipotezy reprezentowanej przez następującą regułę, wykorzystującą wszystkie dostępne atrybuty (pełna zależność funkcjonalna):

$$\begin{aligned} \text{H: JEŻELI } & (\text{SIZE} = \text{b. mały}) \text{ ORAZ } (\text{CYL} = 4) \\ & (\text{TURBO} = \text{nie}) \text{ ORAZ } (\text{FUELSYS} = \text{efi}) \\ & (\text{DISPLACE} = \text{mała}) \text{ ORAZ } (\text{COMP} = \text{wysoka}) \\ & (\text{POWER} = \text{średnia}) \text{ ORAZ } (\text{TRANS} = \text{manual}) \\ & (\text{WEIGHT} = \text{lekka}) \text{ TO } (\text{MILEAGE} = \text{duży}) \end{aligned} \tag{2.13}$$

Wiedza dziedzinowa może mieć następującą postać:

$$\begin{aligned} \text{WD} = \{ & (\text{SIZE} = \text{b. mały}) \Rightarrow (\text{WEIGHT} = \text{lekka}), \\ & (\text{TURBO} = \text{nie}) \Rightarrow (\text{POWER} = \text{średnia})\} \end{aligned} \tag{2.14}$$

Poprzez zastosowanie WD do wstępnej hipotezy, warunki hipotezy: (POWER = średnia) i (WEIGHT = lekka) mogą zostać usunięte z hipotezy. Po tym ocenia się hipotezę na podstawie danych i można usunąć dodatkowe nieistotne warunki z hipotezy podczas odkrywania wiedzy.

Weryfikacja odkrywania potencjalnie sprzecznych reguł. Wiedza dziedzinowa może być także zastosowana do testowania poprawności odkrytej wiedzy. Generalnie, wiedza może być wykorzystana do zweryfikowania, czy odkryta sprzeczna wiedza jest rzeczywiście sprzeczna czy też odkryta, możliwie zgodna wiedza, jest w istocie niepoprawna. Dla przykładu załóżmy, że jesteśmy zainteresowani znalezieniem czynników, które wywołują napad astmy. Eksploracja danych może odkryć następującą wiedzę:

$$\begin{aligned} &\text{Reguła 1: JEŻELI (stan zapalny=tak) ORAZ (wysiłek=tak)} \\ &\qquad\quad \text{TO (napad astmy=tak)} \\ &\text{Reguła 2: JEŻELI (stan zapalny=tak) ORAZ (wysiłek=tak)} \\ &\qquad\quad \text{TO (napad astmy=nie)} \end{aligned} \tag{2.15}$$

Na pierwszy rzut oka wydaje się, że te dwie odkryte reguły są sprzeczne. Jednak mamy dostępną dodatkową wiedzę dziedzinową mówiącą, że pewien lek X zapobiega występowaniu napadów astmy. Zatem wiedza dziedzinowa sprawdza, czy odkryta wiedza jest poprawna czy raczej sprzeczna. Pojawia się więc pytanie, czy można wykorzystać wiedzę dziedzinową do określenia dokładniejszej hipotezy, celem uniknięcia generowania reguł, które wydają się być sprzeczne. Podstawowym pomysłem jest, aby rozwinąć hipotezę dodając więcej warunków na podstawie dostępnej wiedzy dziedzinowej. Należy zbadać zbiór dostępnej wiedzy dziedzinowej i znaleźć każdą regułę, która obejmuje cel do odkrycia. Załóżmy, że dla powyższego przykładu astmy mamy następującą wiedzę dziedzinową:

$$\text{WD} = \{(\text{lek X} = \text{tak}) \Rightarrow (\text{napad astmy} = \text{nie})\} \tag{2.16}$$

Należy zatem dodać informację o stosowaniu leku X do hipotezy. Dzięki temu możemy uzyskać następujące reguły, które nie wydają się być sprzeczne.

$$\begin{aligned} &\text{Reguła 1: JEŻELI (stan zapalny=tak) ORAZ (wysiłek=tak)} \\ &\qquad\quad \text{ORAZ (lek X=nie) TO (napad astmy=tak)} \\ &\text{Reguła 2: JEŻELI (stan zapalny=tak) ORAZ (wysiłek=tak)} \\ &\qquad\quad \text{ORAZ (lek X=tak) TO (napad astmy=nie)} \end{aligned} \tag{2.17}$$

Zapobieganie odkrywaniu ewentualnych reguł nadmiarowych. Zbiory danych często zawierają dane nadmiarowe, które mogą prowadzić do odkrywania zbędnych reguł. Przykładowo dane medyczne dotyczące przewlekłej niewydolności nerek (PNN) mogą zawierać między innymi informacje na temat wzrostu, wagi oraz powierzchni ciała BSA (ang. *body surface area*). BSA służy do wyznaczania przesączania kłębuszkowego GFR (ang. *glomerular filtration rate*), parametru

oceniającego pracę nerek. Nadmiarowy atrybut BSA wyznaczany jest ze wzoru [66]:

$$BSA = 0.15058 \cdot wzrost^{0.3964} \cdot waga^{0.5378} \tag{2.18}$$

Załóżmy, że w procesie odkrywania wiedzy jako cel określono duży stopień nasilenia PNN, gdzie reszta atrybutów stanowi przesłankę. W procesie mogą zostać odkryte reguły wiążące BSA z zaawansowaną PNN, jak i wagę wraz ze wzrostem z zaawansowaną PNN. Chociaż odkryte reguły oparte na BSA oraz na wadze wraz ze wzrostem są różne pod względem składni, to semantycznie są takie same. Nadmiarowe informacje w zbiorze danych można potraktować jako wiedzę dziedzinową i stosować je w procesie odkrywania celem uniknięcia generowania reguł, które różnią się składnią, ale semantycznie są równoważne. Przed etapem eksploracji danych, należy sprawdzić dostępną wiedzę, aby odnaleźć reguły, których atrybuty zawarte są w hipotezie. Jeżeli taka wiedza występuje, to atrybuty tylko jednej strony WD powinny być włączone do procesu odkrywania. W przedstawionym przykładzie PNN, można zastosować tylko BSA lub tylko wagę ze wzrostem. Wybór zależy od tego, czy chcemy wygenerować bardziej ogólne reguły czy też bardziej szczegółowe. Zaletą takiego podejścia jest nie tylko korzyść z zapobiegania generowaniu zbędnych reguł, ale również generowanie reguł, które są bardziej znaczące. Na przykład dla zaawansowanej PNN mogą zostać wygenerowane reguły oparte na BSA i wadze jak i na BSA i wzroście, które nie wydają się być znaczące, ponieważ atrybut wzrost czy waga samodzielnie nie ma powiązania z BSA. Zastosowanie wiedzy dziedzinowej może także zapobiegać odkrywaniu trywialnej wiedzy. Na przykład odkryta reguła: *Im wyższy poziom cukru, tym bardziej nasilona cukrzyca* nie jest niczym odkrywczym, ponieważ jest to znany fakt.

Zapobieganie blokowaniu odkrywania nieoczekiwanej wiedzy. Głównym celem wykorzystania wiedzy dziedzinowej w procesie odkrywania wiedzy jest nastawienie na poszukiwanie ciekawych wzorców poprzez skupianie się na wybranych obszarach danych. Uzyskaną korzyścią jest większa wydajność procesu i bardziej istotne odkrycia. Jednak zbyt duże poleganie na wiedzy dziedzinowej, może ograniczać odkrywanie wiedzy i blokować nieoczekiwane odkrycia np. poprzez niezbadanie części danych. Dla przykładu, załóżmy że chcemy odkryć *Wpływ leku X na pacjentów z chorobą niedokrwienną serca.* Wiedza dziedzinowa sugeruje, że *Osoby poniżej 30 roku życia nie chorują na chorobę niedokrwienną.* Ta wiedza pozwala na zmniejszenie rozmiaru danych poprzez wyeliminowanie przykładów pacjentów w wieku poniżej 30 lat. Załóżmy też, że odkryta wiedza ma postać: *Lek X powoduje efekt A u pacjentów z CNS.* Gdyby nie zastosowało się WD, proces odkrywania wiedzy mógłby znaleźć bardziej rozsądny wynik, taki jak *Lek X powoduje efekt A u pacjentów z CNS powyżej 30 roku życia* oraz *Lek X powoduje efekt B u pacjentów*

z CNS poniżej 30 roku życia. W pewnych przypadkach wykluczenie zastosowania wiedzy dziedzinowej podczas odkrywania może przyczynić się do bardziej efektywnego klasyfikowania danych. Na przykład dane mogą wspierać teorię, że lek X wywołuje różne skutki u osób poniżej 30 roku życia i powyżej 30 roku życia. Jednak ze względu na wyeliminowanie części przykładów dla pacjentów poniżej 30 roku życia, proces odkrywania nie może znaleźć wystarczającej ilości danych na poparcie tej teorii. Podobnie, jeśli używamy WD postaci: *Pacjenci płci męskiej nie chorują na raka piersi* do badania hipotezy: *Wpływ leczenia lekiem X chorych na raka piersi*, można nie dowiedzieć się nieoczekiwanej wiedzy, że pacjenci płci męskiej również mogą zachorować na raka piersi [122]. Należy więc zachować ostrożność w zastosowaniu wiedzy dziedzinowej do zawężania poszukiwań w danych, aby uniknąć zablokowania odkrywania nieoczekiwanej wiedzy. Można to osiągnąć na kilka sposobów. Po pierwsze, ekspert dziedzinowy może przypisać współczynnik ufności CNF każdej regule ze zbioru WD i używać tylko tych reguł, których współczynnik CNF jest większy od określonej wartości progowej. Przypisanie wiedzy dziedzinowej wartości CNF zależy od tego, jak zbieżna jest wiedza dziedzinowa z ustalonymi faktami. Na przykład, biorąc pod uwagę znane fakty, wiedza dziedzinowa postaci: "mężczyźni nie mogą być w ciąży" powinna otrzymać wyższą wartość współczynnika ufności niż wiedza: "kobiety w wieku poniżej 12 i powyżej 65 roku życia nie mogą być w ciąży", ponieważ pierwsza jest niemożliwa z punktu widzenia medycznego, natomiast w drugim przypadku istnieje niewielka szansa, że kobieta poniżej 12 lat lub powyżej 65 lat może zajść w ciążę. Ekspert powinien zdefiniować mechanizm obliczania współczynnika ufności wiedzy dziedzinowej.

Po drugie, rzadko zdarza się, że odkryta wiedza jest prawdziwa dla wszystkich danych. Reprezentowanie i dostarczanie stopnia pewności jest ważne, aby określić w jakim stopniu użytkownik może zaufać wynikom danego procesu odkrywania wiedzy. Pewność ta obejmuje kilka czynników, w tym integralność danych, wielkość próby, na której dokonywane są odkrycia, a także stopień wsparcia ze strony dostępnej wiedzy dziedzinowej. W związku z tym, jeżeli rozmiar zbioru danych jest drastycznie redukowany po zastosowaniu WD, to należy rozważyć użycie tej wiedzy w mniejszym zakresie, albo z niej zrezygnować, w celu uniknięcia blokowania nieoczekiwanych wyników. W przeciwnym razie odkryta wiedza nie ma wystarczająco wysokiego współczynnika ufności, aby uznać ją za interesującą.

Po trzecie, używając wiedzy dziedzinowej w zbyt dużym zakresie można otrzymać wysoce wyspecjalizowany system, być może bardziej efektywny niż jakikolwiek ogólny schemat, jednak nieprzydatny poza konkretną dziedziną. Wiedza dziedzinowa może być wykorzystywana efektywniej poprzez opracowanie ogólnego schematu odkrywania wiedzy, a następnie rozszerzenie go o specyficzną wiedzę dziedzinową [113].

2.6 Dotychczasowe badania nad zastosowaniem wiedzy dziedzinowej

Literatura w dziedzinie odkrywania wiedzy z danych dostarcza wielu przykładów zastosowania wiedzy dziedzinowej w procesie KDD. W podrozdziale przedstawione zostaną podejścia do tego zagadnienia dla procesu KDD oraz osobno w zadaniu klasyfikacji i tworzenia odległości semantycznej, które są tematem rozprawy.

Przegląd istniejących podejść do problemu odkrywania wiedzy z wykorzystaniem wiedzy dziedzinowej

Różni badacze przedstawili sugestie dotyczące roli wiedzy dziedzinowej w KDD. Brachman i Anand [27] zwrócili uwagę, że wiedza dziedzinowa powinna prowadzić proces KDD i nim kierować. Fayyad i wsp. [52] sugerują, że zastosowanie wiedzy dziedzinowej jest ważne we wszystkich etapach procesu odkrywania wiedzy. Domingos [47] sugeruje wykorzystanie tej wiedzy jako najbardziej obiecującego podejścia do zawężania odkrywania wiedzy oraz dla uniknięcia znanego problemu nadmiernego dopasowania odkrytych modeli do zbioru uczącego. Yoon i wsp. [171] proponują następującą klasyfikację wiedzy dziedzinowej: wiedza międzyatrybutowa, która opisuje zależności między atrybutami, wiedza kategorii dziedzinowych, która reprezentuje użyteczne kategorie wartości atrybutów i wiedza korelacji dziedzinowych sugerująca korelacje między atrybutami. W podobny sposób Anand i wsp. [8] identyfikują następujące formy wiedzy dziedzinowej: reguły relacji między atrybutami AR-rules (ang. *Attribute Relationship Rules*), hierarchiczne drzewa uogólniania HG-Trees (ang. *Hierarchical Generalization Trees*) i więzy EBC (ang. *Environment-Based Constraints*). Przykładem więzu jest określenie stopnia zaufania do różnych źródeł danych. Autorzy zastosowali wiedzę dziedzinową w celu zmniejszenia przestrzeni poszukiwań przed fazą eksploracji danych, co dało bardziej intuicyjne wzorce. W innym badaniu, Ambrosino i Buchanan [7] badali, czy dodanie wiedzy dziedzinowej poprawia indukcję reguł w przewidywaniu ryzyka zgonu u pacjentów z pozaszpitalnym zapaleniem płuc. Rozszerzone modele osiągały znacznie lepsze wyniki (niższy średni błąd) niż modele bez wiedzy. Zastosowanie wiedzy dziedzinowej polegało na dodaniu nowych atrybutów, które zostały pozyskane na podstawie istniejących atrybutów. Według Pohle [115] techniki eksploracji danych są skuteczne w generowaniu użytecznych statystyk oraz znajdowaniu wzorców w dużych zbiorach danych, ale nie są tak skuteczne w interpretacji tych wyników, w czym może pomóc wiedza dziedzinowa. Dybowski i wsp. [49] badali, w jaki sposób można łączyć techniki eksploracji danych z wiedzą dziedzinową, aby skonstruować bardziej użyteczne, efektywne i skuteczne systemy wspomagania decyzji. W innym badaniu, Weiss i wsp. [165] połączyli system ekspertowy z metodami eksploracji danych do uzyskania lepszej identyfikacji przyszłych klien-

tów. Opracowali system ekspertowy prowadzący wywiady z menadżerami małych i średnich przedsiębiorstw, który na podstawie ich odpowiedzi, zaleca rozpoznawanie przyszłych klientów. Pary pytanie-odpowiedź i zalecane rozwiązania były przechowywane jako przykłady przeznaczone do eksploracji metodą indukcji reguł. Badanie pokazało, w jaki sposób baza wiedzy może być wykorzystywana do naprowadzania procesu odkrywania wiedzy. Autorzy wskazują, że techniki opracowane w badaniu mogą być przydatne dla systemów konsultacyjnych. Znaczenie ludzkiej inteligencji w eksploracji danych zostało zbadane przez Sharma i Osei-Bryson [133]. Naukowcy zidentyfikowali dwanaście procesów eksploracji danych, które wymagają ludzkiej inteligencji. Uznano, że DM wymaga ludzkiej inteligencji w celu wygenerowania ważnych wyników. Chien i Chen [37] współpracowali z ekspertami dziedzinowymi nad utworzeniem specyficznej procedury rekrutacji pracowników i strategiami zarządzania zasobami ludzkimi z wykorzystaniem technik eksploracji danych. Ich wyniki zostały z powodzeniem zastosowane w rzeczywistej działalności gospodarczej. Singh i Nagpal [137] zaproponowali algorytm IAR (ang. *Interactive Association Rule Mining*), stanowiący modyfikację algorytmu Apriori. W podejściu tym ekspert dziedzinowy wskazuje interesujące go atrybuty. Transakcje niezawierające tych atrybutów są usuwane, co prowadzi do odkrywania tylko reguł z wybranymi atrybutami. W wyniku zastosowania takiego podejścia generowano mniej zbiorów częstych (ang. *frequent itemsets*), uzyskując w wyniku krótszy czas odkrywania reguł. Na podstawie przeglądu przedstawionego przez Cao i wsp. w [35], można stwierdzić, że przeprowadzono wiele innych badań wskazujących na ważne znaczenie wiedzy dziedzinowej w eksploracji danych. Również w publikacjach [13, 16] przedstawiono podejścia mające na celu poprawienie za pomocą wiedzy dziedzinowej jakości klasycznych metod konstruowania klasyfikatorów, takie jak wprowadzanie podpowiedzianych przez eksperta wag klas decyzyjnych czy dyskretyzacja atrybutów wsparta za pomocą wiedzy dziedzinowej.

Przegląd istniejących podejść do klasyfikacji z wykorzystaniem wiedzy dziedzinowej

W jednym z pierwszych badań na ten temat, Pazzani i Kibler [108] opracowali algorytm uczenia ogólnego przeznaczenia o nazwie FOCL (ang. *First Order Combined Learner*), który łączył uczenie oparte na wyjaśnieniach z uczeniem indukcyjnym. W późniejszej pracy, Pazzani i wsp. [109] przeprowadzili eksperyment porównujący FOCL z wiedzą dziedzinową z FOCL bez tej wiedzy. Jako wiedzę dziedzinową zastosowano fragment bazy wiedzy systemu ekspertowego. Autorzy stwierdzili, że włączenie wiedzy dziedzinowej znacząco zmniejsza liczbę błędnych klasyfikacji, gdy powiększa się zbiory treningowe. Hirsh i Noordewier [69] zastosowali wiedzę dziedzinową dotyczącą biologii molekularnej do wyrażania danych za pomocą cech wyższego poziomu. Prowadzili oni eksperymenty z cechami wyższego

poziomu i bez nich, stosując drzewa decyzyjne C4.5 i sztuczne sieci neuronowe z propagacją wsteczną w zadaniach klasyfikacji sekwencji DNA (promotorowych i typu splice-junction). Dane surowe (sekwencje 60 nukleotydów) zastąpiono 19 cechami, takimi jak np. obecność wzorców GTG/CAC (związane z interakcjami DNA-białko), właściwości fizyczne i chemiczne sekwencji (proporcja A i T wpływająca na temperaturę rozwijania helis DNA), kształt helisy DNA (na podstawie pewnej kolejności zasad). Odsetek błędów metody C4.5 wynosił 20.4% dla cech niskopoziomowych i 8.7% dla wysokopoziomowych podczas klasyfikacji promotorów oraz 13.2% dla cech niskopoziomowych i 4.2% dla wysoko-poziomowych podczas klasyfikacji splice-junctions. Podobne wyniki uzyskano dla sieci neuronowych. W przypadku obu metod, wykorzystanie cech wyższego poziomu dało w wyniku znacząco niższe wskaźniki błędów. Ciekawe, że połączenie obu typów cech dało gorsze wyniki niż dla samych tylko cech wysokiego poziomu: 10.6% i 5.1%, odpowiednio dla promotorów i splice-junction. Ware i wsp. zaproponowali w [163] interaktywne podejście do budowy drzew decyzyjnych, w których podziały węzłów wskazywane są przez użytkownika (eksperta). System na każdym etapie budowy drzewa umożliwia wizualizację danych w węzłach i aktualizuje je stosownie do wybranych podziałów. Wyniki w postaci dokładności klasyfikacji ACC (patrz Rozdz. 3.3), uzyskane za pomocą takiego podejścia dla zbioru Iris (z repozytorium [158]), powszechnie stosowanego do testowania metod klasyfikacji, były porównywalne z wynikami metody C4.5, ale uzyskiwano mniejsze rozmiary drzew. Zastosowanie abstrakcyjnych atrybutów do budowy drzew decyzyjnych zaproponowali Zhang i wsp. w [176]. W podejściu tym wykorzystywana jest hierarchiczna taksonomia wartości atrybutów AVT (ang. *attribute value taxonomy*), w której na wyższych poziomach znajdują się atrybuty abstrakcyjne, stanowiące zgrupowane wartości niższego poziomu (np. sok jabłkowy i sok pomarańczowy reprezentują abstrakcyjny atrybut: sok owocowy). Algorytm wybiera nie tylko atrybut, ale także jego poziom abstrakcji. Rola eksperta polega na utworzeniu taksonomii atrybutów stanowiącej sposób reprezentacji wiedzy dziedzinowej. Wykorzystanie tego podejścia np. do danych dotyczących nowotworów złośliwych piersi dało zmniejszenie odsetka błędów klasyfikacji metody C4.5 z 34% do 29% [175]. W badaniach przeprowadzonych przez Sinha i Zhao [138] zastosowano wiedzę dziedzinową przy wyznaczaniu zdolności kredytowej z wykorzystaniem 7 metod klasyfikacji (naiwny klasyfikator Bayesa, regresja logistyczna, drzewa i tablice decyzyjne, sztuczne sieci neuronowe, metoda k-najbliższych sąsiadów i SVM. Wiedza dziedzinowa miała postać dodatkowego atrybutu wyznaczonego na podstawie reguł eksperta (ocena możliwości spłaty kredytu w zakresie 0-100%). Ustalono, że istnieje zależność pomiędzy sposobem klasyfikacji a wiedzą dziedzinową. Badacze stwierdzili, że zastosowanie wiedzy dziedzinowej ma większy wpływ na wydajność niektórych metod eksploracji danych, niż innych. Z wyjątkiem drzew decyzyjnych uzyskano statystycznie

istotną poprawę jakości klasyfikacji w porównaniu z modelami bez wiedzy dziedzinowej. Zhao i in. [177] zbadali wpływ zastosowania wiedzy dziedzinowej na wyniki przewidywania upadłości banków. W oparciu o wiedzę dziedzinową skonstruowano 26 nowych zmiennych (wskaźniki finansowe), których zastosowanie w 4 badanych metodach klasyfikacji (regresja logistyczna, drzewa decyzyjne, sztuczne sieci neuronowe i metoda k-najbliższych sąsiadów) dało statystycznie istotną poprawę jakości klasyfikacji. Wyniki ich badań wykazały, że takie podejście znacznie poprawia wydajność klasyfikatora, przy czym była ona najmniejsza dla drzew decyzyjnych. Podejście zaproponowane przez Redouane i wsp. w [124] polega na podziale zbioru danych przez eksperta dziedzinowego na niezależne semantycznie części. Każda część jest traktowana jako szum informacyjny dla pozostałych części zbioru i dlatego dla każdej z nich budowane jest osobne drzewo decyzyjne. Podejście to zastosowano do klasyfikacji chorób gruczołu tarczowego, oddzielając infekcje tego narządu występujące u dorosłych od tych występujących u dzieci. Uzyskane w ten sposób średnie ACC dla podzbiorów wynosiło 68%, natomiast dla całego zbioru: 66%.

Przegląd istniejących podejść do konstrukcji odległości z wykorzystaniem wiedzy dziedzinowej

Przegląd istniejących podejść do wyznaczania odległości między pojęciami przedstawiono w [110] oraz [152]. Miary podobieństwa semantycznego i pokrewieństwa podzielono tam na takie rodzaje jak: oparte na ścieżkach w ontologii pojęć, oparte na zawartości informacji oraz na wektorach kontekstowych. Rada i wsp. [121] definiują pojęcie odległości semantycznej jako długość najkrótszej ścieżki łączącej dwa pojęcia w ontologii pojęć. Im dłuższa ścieżka, tym bardziej oddalone semantycznie są pojęcia. Miarę podobieństwa semantycznego pomiędzy pojęciami opartą na długości oraz głębokości ścieżki zaproponowali Wu i Palmer w [170]. Podejście to wykorzystuje liczbę krawędzi typu `'is-a'` od pojęć do najbliższego wspólnego przodka LCS (ang. *lowest common subsumer*) oraz liczbę krawędzi do korzenia taksonomii. Leacock i Chodorow [86] zaproponowali miarę podobieństwa semantycznego opartą na najkrótszej ścieżce w leksykalnej bazie danych WordNet [169]. Długość ścieżki jest skalowana z wykorzystaniem maksymalnej głębokości taksonomii do wartości z przedziału od 0 do 1, a podobieństwo jest wyliczane jako ujemny logarytm z tej wartości. Miara podobieństwa oparta na pojęciu zawartości informacji IC (ang. *Information Content*) została przedstawiona przez Resnika w [125]. IC będąca miernikiem specyficzności pojęcia, jest obliczana dla każdego pojęcia w hierarchii na podstawie częstości występowania tego pojęcia w szerszym kontekście. Wykorzystując pojęcie IC, Resnik proponuje miarę, w której podobieństwo semantyczne dwóch pojęć jest proporcjonalne do ilości informacji, którą dzielą. Lin w [88] zaproponował rozszerzenie pracy Resnika, poprzez skalowanie

zawartości informacyjnej pojęcia nadrzędnego LCS przez zawartość informacyjną poszczególnych pojęć. Hsu i wsp. [70] przedstawili reprezentację odległości w postaci hierarchii odległości stanowiącej rozszerzenie hierarchii pojęć poprzez nadanie wag połączeniom. Odległość między dwiema wartościami atrybutu (kategorycznego lub numerycznego) jest mierzona jako całkowita waga połączeń na ścieżce między dwoma węzłami pojęć. Wagi określane są przez eksperta (wiedza dziedzinowa). Proponowane podejście zastosowane w algorytmie grupowania z użyciem hierarchicznej metody aglomeracyjnej lepiej ukazywało podobieństwo struktury danych.

Wymienione metody wyznaczania odległości semantycznej z wykorzystaniem WD dotyczą budowy odległości między pojęciami lub wartościami atrybutów, przez co znajdują zastosowanie np. w dyskretyzacji atrybutów. Proponowana natomiast w rozprawie metoda konstrukcji odległości ontologicznej stanowi odmienne podejście, mające na celu porównywanie podobieństwa między obiektami przynależącymi do pojęć, podobnie jak w metodzie mierzenia podobieństwa pomiędzy planami we wcześniejszej pracy autorki rozprawy [16].

Jak pokazano w przedstawionym przeglądzie literatury na temat zastosowania WD w procesie KDD, wiedza dziedzinowa jest stosowana w różnym stopniu na wielu etapach i w zróżnicowanej postaci. Trzeba także zwrócić uwagę na fakt, że wiele obecnych narzędzi do odkrywania wiedzy nie umożliwia reprezentowania wiedzy dziedzinowej. Wykorzystanie tej wiedzy w praktyce jest najczęściej obsługiwane ręcznie, poprzez wyeliminowanie np. zbędnych atrybutów dla konkretnego problemu decyzyjnego [103].

2.6.1 Problemy we wdrażaniu wiedzy dziedzinowej do procesu odkrywania wiedzy

Pomimo wielu doniesień o możliwościach poprawiania efektywności odkrywania wiedzy z danych z wykorzystaniem wiedzy dziedzinowej, wciąż nie opracowano jednolitej metodologii jej zastosowania. Wynika to z wielu przyczyn, wśród których można wyróżnić trudności z dostępem do ekspertów, pozyskiwaniem wiedzy od ekspertów, jej reprezentacją i różne obszary zastosowań. Nabywanie wiedzy jest zwykle trudnym i czasochłonnym zadaniem [164], ponieważ eksperci często nie potrafią wyrazić heurystyk lub zasad, które służą im do skutecznego rozwiązania problemów decyzyjnych. To zjawisko nazywane jest *wąskim gardłem* w procesie akwizycji wiedzy (ang. *knowledge acquisition bottleneck*) [138]. Co więcej, im bardziej kompetentnym staje się ekspert, tym mniej jest on w stanie opisać wykorzystywaną przez siebie wiedzę do rozwiązywania problemów [74]. Istotny jest także ograniczony czas, który ekspert może poświęcić na opisywanie swojej wiedzy. Wśród metod mających na celu rozwiązanie przedstawionych problemów w pozyskiwaniu wiedzy opracowano na gruncie inżynierii wiedzy wiele technik ułatwiających to zadanie. Należą do nich m.in.: przeprowadzanie wywiadów, protokoły analiz czy obserwacje. Zwraca się uwagę także na odpowiedni dobór eksperta, od którego wiedza będzie pozyskiwana. Wybór powinien opierać się na osiągnięciach i doświadczeniu eksperta, a także łatwości komunikacji z inżynierem wiedzy. Niestety, wiedza dziedzinowa często jest nieformalna i trudno strukturyzowalna. Trudno jest zatem wcielać tę wiedzę do standardowych metod eksploracji danych.

Z drugiej strony należy zachować pewną ostrożność przy stosowaniu wiedzy dziedzinowej w odkrywaniu wiedzy z danych, o czym już wspomniano w tym rozdziale.

Rozdział 3

Wybrane metody tworzenia klasyfikatorów

Zawartość

Metody budowy klasyfikatorów próbują odkryć zależność między zmiennymi objaśniającymi (predykcyjnymi) oraz zmienną celu. Odkryty związek jest zawarty w strukturze zwanej modelem. Zazwyczaj modele opisują i wyjaśniają zjawiska ukryte w zbiorze danych i mogą być używane do przewidywania wartości atrybutu decyzyjnego na podstawie wartości atrybutów warunkowych. Zadanie klasyfikacji polega zatem na modelowaniu granic między klasami. Utworzone modele dokonują podziału całej przestrzeni na obszary odpowiadające klasom.

Do utworzenia modelu predykcyjnego potrzebny jest zbiór obiektów, dla których znane są wartości zarówno zmiennych predykcyjnych jak i zmiennych celu (zbiór uczący), aby można było przewidywać wartość zmiennej celu dla nowych obiektów, dla których znane są tylko wartości zmiennych predykcyjnych. Konstrukcja modelu może odbywać się poprzez znalezienie parametrów funkcji separującej

(sieci neuronowe, metoda wektorów podpierających), wygenerowanie zestawu reguł bądź drzew decyzyjnych, czy też znalezieniu parametrów rozkładu (regresja). Cały proces klasyfikacji można podzielić na dwa etapy [63]:

1. Uczenia - konstrukcja modelu w oparciu o zbiór danych (przykłady uczące).
2. Klasyfikacji - zastosowanie modelu do predykcji etykiet klas dla nowych obiektów.

Na początku etapu drugiego ocenia się dokładność predykcji modelu (klasyfikatora). Jeżeli dokładność predykcji jest akceptowalna, model może zostać wykorzystany do klasyfikacji przyszłych (nowych) danych, dla których wartość atrybutu decyzyjnego jest nieznana.

Na przestrzeni ostatnich kilkudziesięciu lat opracowano wiele algorytmów klasyfikacji (patrz [95, 64, 65, 92, 63, 174]), różniących się takimi właściwościami jak: jakość klasyfikacji, szybkość klasyfikacji, szybkość uczenia, zapotrzebowania pamięciowe czy złożoność obliczeniowa, z których najczęściej wymienia się takie metody jak:

- Dyskryminacja liniowa, LDA (ang. *Linear Discriminant Analysis*) i kwadratowa, QDA (ang. *Quadratic Discriminant Analysis*) [95, 65, 174] - polegają na znalezieniu liniowej czy kwadratowej kombinacji cech, określających granice między klasami;
- Metoda najbliższych sąsiadów, k-NN (ang. *k-Nearest Neighbours*) [95, 63, 174] - nowemu obiektowi przypisuje się klasę, która występuje najczęściej pośród jego k sąsiadów. Sąsiedztwo oznacza k najbliższych obiektów znajdujących się w zbiorze uczącym, najbliższych w sensie określonej miary odległości;
- Naiwny klasyfikator Bayesa, NB (ang. *Naive Bayes*) [92, 63, 168, 174] - jedna z metod bazujących na ocenie prawdopodobieństwa przynależności do określonej grupy. Opiera się na regule Bayesa użytej do wyznaczenia prawdopodobieństwa a posteriori należenia do poszczególnych klas;
- Drzewa decyzyjne, DT (ang. *Decision Trees*) [95, 65, 92, 168, 174] - klasyfikator jest reprezentowany przez drzewo, w którego węzłach znajdują się pytania o wartości określonej cechy, a w liściach oceny klas. Aby zbudować drzewo klasyfikacyjne należy określić kryterium podziału oraz kryterium zatrzymania podziału (stopu);
- Sztuczne sieci neuronowe, ANN (ang. *Artificial Neural Networks*) [95, 92, 63] - systemy, których struktura jest wzorowana na działaniu ludzkiego układu nerwowego, realizujące obliczenia poprzez rzędy elementów, zwanych sztucznymi neuronami. SNN uczą się zadanej funkcji poprzez obserwowanie przy-

kładów jej działania, a proces uczenia polega na modyfikowaniu wag neuronów;

- Metoda wektorów nośnych SVM (ang. *Support Vector Machines*) [92, 63, 174] - ideą wykorzystywaną w tej technice jest transformacja zmiennych oryginalnych, tak aby obiekty różnych klas można było rozdzielić hiperpłaszczyznami i wybór spośród wielu możliwych tego typu rozwiązań, optymalnych w określonym sensie.

Ponieważ w rozprawie do sprawdzenia prawdziwości tez zastosowano dwie techniki tworzenia klasyfikatorów: drzewa decyzyjne oraz metodę k-NN, zostaną one omówione dokładniej.

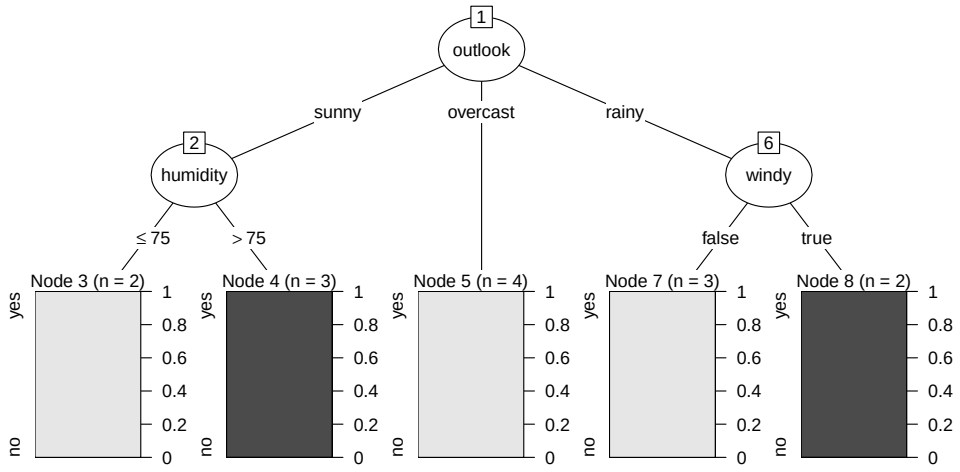
3.1 Drzewa decyzyjne

Drzewo decyzyjne to struktura drzewiasta w sensie teorii grafów reprezentująca proces podziału zbioru obiektów na klasy. Węzły wewnętrzne opisują sposób dokonania tego podziału, liście odpowiadają klasom, do których przynależą obiekty. Natomiast krawędzie drzewa reprezentują wartości cech, na podstawie których dokonano podziału [119]. W oryginalnej koncepcji drzew decyzyjnych [119, 120, 30] jako kryterium wyboru testu podczas budowy drzewa stosowane są takie miary jak: entropia, zysk informacji czy współczynnik korzyści. Drzewo decyzyjne buduje się w sposób rekurencyjny od korzenia do liścia w oparciu o zasadę "dziel i rządź" (ang. *divide and conquer*), która polega na podziale złożonego problemu na prostsze podzadania, a następnie rekursywnym zastosowaniu tej strategii do utworzonych podzadań. Klasyfikacja odbywa się poprzez dopasowywanie klasyfikowanego obiektu do ścieżki od korzenia do liścia zgodnie z wynikami testów.

Przykład drzewa decyzyjnego utworzonego dla zbioru danych *WeatherPlay* opisującego warunki pogodowe przedstawia Rys. 3.1. W zbiorze tym dostępne są dwa atrybuty numeryczne: temperatura (ang. *temperature*) oraz wilgotność powietrza (ang. *humidity*), jeden symboliczny wielowartościowy: warunki zewnętrzne (ang. *outlook*), jeden symboliczny dwuwartościowy: wietrzna pogoda (ang. *windy*) oraz atrybut granie (ang. *play*) stanowiący zmienną decyzyjną. Celem analizy jest określenie, czy w daną pogodę dobrze gra się w golfa. Drzewo posiada 3 węzły wewnętrzne oraz 5 węzłów końcowych (liści). W każdym liściu na rysunku podano liczbę obiektów tworzących węzeł oraz proporcję klas decyzyjnych. Liście przykładowego drzewa zawierają obiekty tylko jednej klasy decyzyjnej, dlatego przedstawione są za pomocą jednego koloru każdy (jasnoszare słupki reprezentują klasę *tak* (ang. *yes*), ciemnoszere - klasę *nie* (ang. *no*).

```
> WeatherPlay
```

```
outlook temperature humidity windy play
1     sunny          85       85 false   no
2     sunny          80       90  true   no
3  overcast          83       86 false  yes
4     rainy          70       96 false  yes
5     rainy          68       80 false  yes
6     rainy          65       70  true   no
7  overcast          64       65  true  yes
8     sunny          72       95 false   no
9     sunny          69       70 false  yes
10    rainy          75       80 false  yes
11    sunny          75       70  true  yes
12 overcast          72       90  true  yes
13 overcast          81       75 false  yes
14    rainy          71       91  true   no
```



Rysunek 3.1: Przykładowe drzewo decyzyjne dla zbioru *WeatherPlay* [156], w którym problem decyzyjny polega na przewidywaniu dobrych warunków pogodowych do gry w golfa.

Zauważmy, że powyższe drzewo decyzyjne można traktować wprost jako klasyfikator, gdyż obiekty testowe mogą być klasyfikowane poprzez stwierdzenie do jakiego liścia drzewa należą. Jest to możliwe, bo dzięki wyznaczonym cięciom można prześledzić przynależność obiektu na ścieżce od korzenia do liścia, po czym sklasyfikować obiekt do klasy decyzyjnej, której obiekty dominują w tym liściu.

Najpopularniejszymi algorytmami na bazie drzew decyzyjnych są "dychotomizer interaktywny" ID3 (ang. *Interactive Dichotomizer, version 3*), CART (ang. *Classification and Regression Trees*) oraz C4.5 [30, 119, 120]. W trzech pierwszych proponowanych w rozprawie metodach wykorzystano klasyfikator oparty na binarnym drzewie decyzyjnym lokalnej dyskretyzacji (patrz np. [101]).

3.1.1 Cięcia i wzorce

Metoda wyboru atrybutu oraz jego wartości, czyli cięcia, które wykorzystywane są do podziału zbioru obiektów, stanowi kluczowy element konstrukcji drzewa decyzyjnego lokalnej dyskretyzacji. Wybór ten powinien uwzględniać badanie wartości atrybutu decyzyjnego obiektów ze zbioru uczącego.

Formalnie, cięcie to para (a, v) zdefiniowana dla danej tablicy decyzyjnej $\mathbf{A} = (U, A, \cup\{d\})$ w sensie zbiorów przybliżonych Pawlaka (patrz [105, 107]), gdzie $a \in A$ (A to zbiór atrybutów lub kolumn w zbiorze danych), natomiast v stanowi wartość atrybutu a.

Dowolne cięcie $c = (a, v)$ definiuje dwa wzorce, gdzie wzorzec oznacza opis zbioru obiektów. Wzorce te określane są odmiennie dla atrybutów numerycznych i symbolicznych. W przypadku atrybutów numerycznych, pierwszy wzorzec na bazie cięcia (a, v), nazywany będzie *lewym wzorcem* i jest formułą: $TL(c) = \{u \in U : a(u) < v\}$, natomiast drugi wzorzec jest formułą: $TR(c) = \{u \in U : a(u) \geq v\}$ i nazywany będzie *prawym wzorcem* .

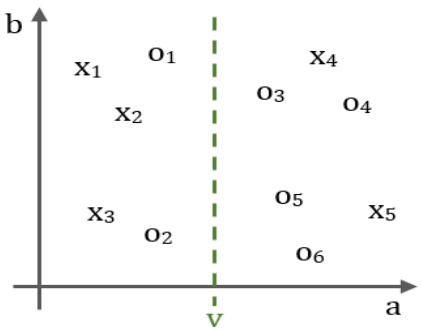
Niech dla danego obiektu $u \in U$ spełniona jest nierówność $a(u) < v$. Wówczas będziemy mówić, że obiekt u pasuje do wzorca (ang. *matches pattern*) lub wspiera wzorzec (ang. *supports pattern*). W przeciwnym razie, obiekt nie pasuje do wzorca $TL(c)$. Analogicznie, obiekt u pasuje do wzorca $TR(c)$, jeżeli prawdziwe jest wyrażenie: $a(u) \geq v$, co oznacza, że wartość atrybutu a obiektu u jest większa lub równa v, w przeciwnym przypadku, obiekt nie pasuje do wzorca $TR(c)$.

Dla atrybutów symbolicznych *lewy wzorzec* jest wzorem: $TL(c) = \{u \in U : a(u) = v\}$, natomiast *prawym wzorcem* jest formuła: $TR(c) = \{u \in U : a(u) \neq v\}$. Obiekt $u \in U$ pasuje do wzorca $TL(c)$, jeżeli spełniona jest równość: $a(u) = v$, czyli wartość atrybutu $a \in A$ dla tego obiektu jest równa v, w przeciwnym wypadku obiekt nie pasuje do wzorca $TL(c)$. W końcu, obiekt $u \in U$ pasuje do wzorca $TR(c)$, gdy pasuje do opisu: $a(u) \neq v$, to znaczy wartość atrybutu a nie równa się v, inaczej obiekt nie pasuje do wzorca.

Jeżeli c jest cięciem, wzorzec zdefiniowany dla c ogólnie oznaczany będzie jako $T(c)$, przy czym może on odpowiadać wzorcowi $TL(c)$ lub $TR(c)$. Dla uproszczenia opisu, w miejsce $T(c)$ stosowany będzie zapis T dla ustalonego już cięcia c. Ponadto, jeżeli T stanowi dowolny spośród dwóch wzorców określony na podstawie cięcia c, wówczas $\neg T(c)$ oznacza wzorzec inny niż T, dla cięcia c. Jeżeli ponadto

zdefiniowany jest wzorzec T dla danej tablicy decyzyjnej $\mathbf{A} = (U, A, \cup\{d\})$, wówczas $\mathbf{A}(T)$ oznacza zbiór wszystkich obiektów z U pasujących do wzorca T.

Para obiektów $(u_1, u_2) \in U \times U$ jest rozróżniana przez cięcie c definiujące wzorzec T, jeżeli u_1 pasuje do wzorca T, natomiast u_2 do niego nie pasuje. Lub odwrotnie, u_2 pasuje do wzorca T, a u_1 nie. Na przykład, pary (x_1, x_2) czy (o_3, o_4) z Rys. 3.2 nie są rozróżniane przez cięcie $c = (a, v)$ definiujące wzorzec T, natomiast pary (x_2, x_4) i (o_1, x_5) są rozróżniane przez c. Przez $Disc(c)$ oznaczana



Rysunek 3.2: Wizualizacja cięcia w przestrzeni dwuwymiarowej.

będzie liczba par obiektów z różnych klas decyzyjnych rozróżnianych przez cięcie c. Sposób wyliczania $Disc(c)$ przedstawiony zostanie na przykładzie obiektów z Rys. 3.2. Na rysunku znajduje się 11 obiektów, przynależących do dwóch klas decyzyjnych: X i O. Do pierwszej klasy należą obiekty: x_1, x_2, x_3, x_4, x_5, do drugiej: $o_1, o_2, o_3, o_4, o_5, o_6$. Wzorzec T zdefiniowany przez cięcie (a, v) dzieli obiekty na następujące dwa podzbiory: $\{x_1, x_2, x_3, o_1, o_2\}$ oraz $\{x_4, x_5, o_3, o_4, o_5, o_6\}$, z których jeden zawiera przykłady pasujące do T, a drugi niepasujące do T. Pierwszy podzbiór zawiera 3 obiekty z klasy X i 2 z klasy O, natomiast drugi: 2 obiekty z klasy X oraz 4 z klasy O. Liczba obiektów rozróżnianych przez cięcie $c = (a, v)$ wynosi zatem: $Disc(c) = 3 \cdot 4 + 2 \cdot 2 = 16$. Po wyznaczeniu wartości tej miary dla wszystkich możliwych cięć, można zachłannie wybrać jedno z cięć i na jego podstawie podzielić zbiór wszystkich obiektów na dwie części. Takie podejście może być z łatwością uogólnione do przypadku z więcej niż dwiema klasami decyzyjnymi. Liczba $Disc(c)$ będzie tutaj traktowana jako miara jakości cięcia c.

Duże znaczenie ma fakt, że powyższa miara jakości cięcia $Disc(c)$ może być wyliczona dla danego cięcia w czasie $O(n)$, gdzie n oznacza liczbę obiektów w tabeli decyzyjnej [23]. Jednak wyznaczenie optymalnego cięcia wymaga wyliczenia miary jakości dla wszystkich potencjalnych cięć. W tym celu należy sprawdzić wszystkie potencjalne cięcia, uwzględniając wszystkie atrybuty warunkowe. Może to być zrealizowane za pomocą wielu sposobów. Jedna z takich metod, w przypadku atrybutów numerycznych, najpierw sortuje wartości danego atrybutu, dla

którego poszukujemy optymalnego podziału. To pozwala na wyznaczenie optymalnego cięcia w czasie liniowym.

Sortowanie wartości atrybutu skutkuje faktem, że wyliczenie optymalnego podziału odbywa się w czasie $O(n \cdot \log n \cdot m)$, gdzie n oznacza liczbę obiektów, a m liczbę atrybutów warunkowych.

3.1.2 Miary jakości podziałów w drzewie decyzyjnym

Na każdym etapie tworzenia drzewa algorytm wybiera zachłannie najlepsze cięcie zgodnie z przyjętą miarą jakości. W rozprawie do budowy drzewa lokalnej dyskretyzacji zastosowano takie miary jak: miara oparta na liczbie par obiektów należących do różnych klas decyzyjnych rozróżnianych przez cięcie, nazywana dalej *DiscPairs*, zysk informacji (ang. *Information Gain*) czy indeks Giniego (ang. *Gini index*). Klasyfikator skonstruowany za pomocą drzewa lokalnej dyskretyzacji i dowolnej z tych miar będzie nazywany dalej klasycznym drzewem decyzyjnym i oznaczany przez CTree *CTree*.

Miara DiscPairs

Jakość cięcia c w zbiorze obiektów X z wykorzystaniem tej miary wyliczana jest według wzoru (3.1):

$$Q_{Disc}(c, X) = \sum_{i,j \in D} M_i \cdot N_j, \quad dla \quad i \neq j \tag{3.1}$$

gdzie D to zbiór klas decyzyjnych, M_i i N_j to liczba obiektów w lewym oraz prawym poddrzewie należących do różnych klas decyzyjnych. Na przykład, niech cięcie c dzieli zbiór obiektów należących do dwóch klas decyzyjnych na dwie grupy o liczebności odpowiednio M i N, a liczba obiektów przynależących do klas C_0 i C_1 niech wynosi M_0 i M_1 w jednej grupie oraz N_0 i N_1 w drugiej. Wówczas liczba par obiektów rozróżnianych przez to cięcie jest dana wzorem (3.2):

$$Q_{Disc}(c, X) = M_0 N_1 + M_1 N_0 \tag{3.2}$$

Jeśli wyznaczymy wartość tej miary dla wszystkich możliwych cięć, to możemy zachłannie wybrać jedno z cięć o najwyższej wartości miary i podzielić cały zbiór obiektów na dwie części na jego podstawie.

Zysk informacji

Miara ta została opisana przy budowie drzew metodą C4.5 [120]. Podejście to wykorzystuje pojęcie entropii opisane przez C. Shannona w jego pracy na temat teorii

informacji [132]. W odniesieniu do konstrukcji drzew decyzyjnych miara ta reprezentuje różnorodność zbioru obiektów, która odpowiada danemu węzłowi w drzewie. Niech ponownie X będzie zbiorem obiektów, który składa się z dwóch klas decyzyjnych: C_0 i C_1. Ponadto, $p_0 = \frac{|C_0|}{|X|}$ i $p_1 = \frac{|C_1|}{|X|}$ są rozkładem C_0 i C_1 w zbiorze X. Wówczas entropia zbioru X jest obliczana za pomocą następującego wyrażenia:

$$Entropia(X) = -\sum_{i=0}^{1} p_i \cdot \log_2 p_i \tag{3.3}$$

Jakość binarnego podziału, który jest określony przez wartość cięcia c w zbiorze X jest obliczana na podstawie miary nazywanej zyskiem informacji w następujący sposób:

$$Q_{Entropy}(c, X) = Entropia(X) - \sum_{i=0}^{1} \frac{|X_i|}{|X|} \cdot Entropia(X_i) \tag{3.4}$$

gdzie X_i dla $i \in \{0, 1\}$ stanowią podzbiory X, które odpowiadają podziałowi zdefiniowanemu przez wartość cięcia c.

Wartość zysku informacji jest określana dla wszystkich możliwych cięć, a następnie zachłannie wybierane jest to cięcie, które maksymalizuje wartość miary. Oczywiście ten przykład można uogólnić na większą liczbę klas decyzyjnych niż dwie.

Indeks Giniego

Jest to miara jakości cięć zastosowana w algorytmie CART [30]. Stosując oznaczenia takie jak przy opisie poprzednich miar, niech X zawiera przykłady z klasy C_0 i C_1. Wówczas miara różnorodności zbioru X jest zdefiniowana jako:

$$Gini(X) = 1 - \sum_{i=0}^{1} p_i^2 \tag{3.5}$$

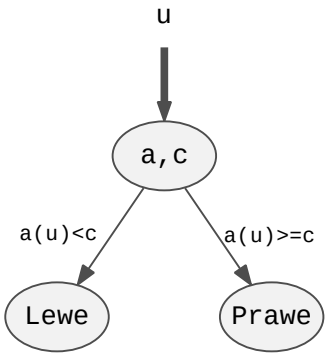
gdzie p_i jest rozkładem klas w X. Jakość cięcia c może być obliczona w następujący sposób:

$$Q_{Gini}(c, X) = Gini(X) - \sum_{i=0}^{1} \frac{|X_i|}{|X|} \cdot Gini(X_i) \tag{3.6}$$

Tak jak wcześniej, najlepszy podział jest wybierany zachłannie ze wszystkich możliwych cięć. Analogicznie jak poprzednio, takie podejście może być uogólnione do więcej niż dwóch klas decyzyjnych.

3.1.3 Budowa drzewa decyzyjnego

W strategii dyskretyzacji lokalnej, po znalezieniu najlepszego cięcia i podzieleniu zbioru obiektów na dwie grupy, procedura ta jest powtarzana rekurencyjnie dla każdego zbioru obiektów, aż do spełnienia warunku zatrzymania. Warto zauważyć, że dyskretyzacja formalnie nie obejmuje grupowania wartości symbolicznych, ale rozważania można przenieść także na przypadek takich wartości. Warunek stopu tworzenia podziałów jest tak skonstruowany, że dana część nie jest już dzielona (staje się liściem drzewa), gdy zawiera obiekty należące do jednej klasy decyzyjnej (alternatywnie, obiekty jednej klasy stanowią określony odsetek, który stanowi parametr metody) lub gdy dalsze podziały nie dają poprawy jakości podziału. Zatem strategia lokalnej dyskretyzacji możne być realizowana w postaci *drzewa decyzyjnego* (patrz Rys. 3.3).



Rysunek 3.3: Drzewo decyzyjne stosowane w lokalnej dyskretyzacji.

W praktyce często dane są zaszumione, co powoduje rozrost drzewa decyzyjnego, utrudnia zrozumienie jego reguł, a także naraża na przeuczenie. Kosztem utraty zgodności (ang. *consistency*), tj. poprawnego rozpoznania absolutnie wszystkich obiektów zbioru uczącego, dąży się do poprawy jakości klasyfikacji nowych obiektów poprzez np. przycinanie (ang. *pruning*) drzewa. Przycinanie polega na zastąpieniu odpowiednio głęboko położonych wierzchołków liśćmi. Wyróżnia się dwie grupy metod przycinania, tj. metody zatrzymujące w odpowiednim momencie tworzenie drzewa (ang. *pre-pruning*) oraz metody generujące pełne drzewo, a następnie dokonujące jego przycięcia (ang. *post-pruning*).

3.1.4 Drzewo decyzyjne jako klasyfikator

Drzewo decyzyjne może być traktowane jako klasyfikator pojęcia C reprezentowanego przez atrybut decyzyjny danej tablicy decyzyjnej DT. Niech u będzie nowym obiektem, a $DT(T)$ podtablicą zawierająca wszystkie obiekty pasujące do wzorca

T zdefiniowanego przez cięcie w bieżącym węźle danego drzewa decyzyjnego. Klasyfikacja obiektu u przebiega według Algorytmu 3.1.1.

Algorytm 3.1.1: Klasyfikacja za pomocą drzewa decyzyjnego

WEJŚCIE: Drzewo decyzyjne, klasyfikowany obiekt u
WYJŚCIE: Przewidywana klasa obiektu u

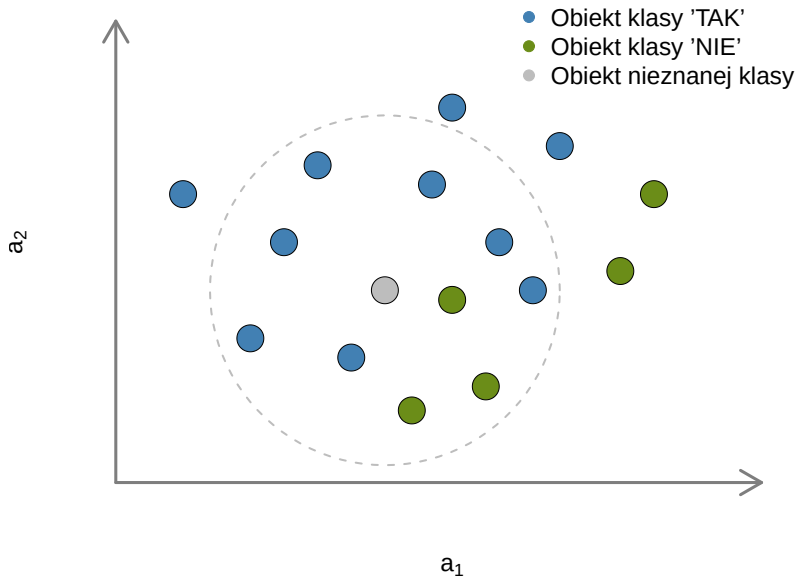
```
1  begin
2      jeżeli u pasuje do wzorca T tablicy DT to
3          idź do poddrzewa związanego ze wzorcem DT(T)
4      inaczej
5          idź do poddrzewa związanego ze wzorcem DT(¬T)
6      jeżeli u znajduje się w liściu to
7          idź do kroku 10
8      inaczej
9          powtórz 2-6 wstawiając DT(T) lub DT(¬T) w miejsce DT
10     Sklasyfikuj u zgodnie z wartością decyzji w liściu
```

Działanie algorytmu rozpoczyna się od węzła zwanego korzeniem (ang. *root*) reprezentującego całą tablicę decyzyjną DT. Sprawdzana jest tutaj zgodność obiektu u ze wzorcem T wyznaczonym dla tego węzła. Jeżeli u pasuje do wzorca T, będącego formułą $\{u \in U : a(u) < v\}$, a więc dla obiektu u zachodzi nierówność $a(u) < v$, wówczas algorytm przechodzi do poddrzewa związanego ze wzorcem $DT(T)$, w przeciwnym razie do poddrzewa związanego ze wzorcem $DT(\neg T)$. Drzewo związane z danym wzorcem zawiera wszystkie obiekty, które do tego wzorca pasują. Następnie sprawdza się, czy węzeł, w którym znajduje się obecnie obiekt u (korzeń danego poddrzewa) jest liściem. Jeżeli tak, wówczas obiekt u otrzymuje etykietę klasy przypisanej do tego węzła podczas budowy drzewa i algorytm kończy swoje działanie. W przeciwnym razie, ponownie sprawdzana jest zgodność u ze wzorcem T wyznaczonym dla obecnego węzła. Procedura jest powtarzana rekurencyjnie dla każdego węzła potomnego i kończy się, gdy węzeł, do którego trafi obiekt u jest liściem. W liściu obiekt u zostaje sklasyfikowany do klasy, która została przypisana do liścia na etapie budowy drzewa.

Klasyfikator zbudowany z wykorzystaniem miary *DiscPairs* nazwany jest tu klasyfikatorem *CTree-Disc*. Klasyfikatory, w których zastosowano miarę opartą na zysku informacji lub indeksie Giniego, to klasyfikator *CTree-Entropy* i *CTree-Gini*, odpowiednio.

3.2 Klasyfikator k-NN

W metodzie k najbliższych sąsiadów, predykcja przynależności nowego obiektu do klasy opiera się na porównaniu go ze zbiorem przykładowych obiektów (patrz np. [95]). O klasyfikacji decyduje głosowanie najbliższych klasyfikowanemu k obiektów (patrz Rys. 3.4). Wymagane jest więc zdefiniowanie funkcji odległości pomiędzy obiektami oraz wybór metody głosowania (zwykle zasada większościowa). W kla-



Rysunek 3.4: Głosowanie podczas klasyfikacji metodą k-NN dla k=10.

sycznych technikach opartych na odległości, stosuje się takie miary jak odległość Minkowskiego (*p-norma*), odległość Euklidesa (*norma L2*) czy miejska (Manhattan, city-block, *norma L1*) [63, 168, 46] wyrażone wzorami 3.7, 3.8 i 3.9.

$$d_{Euclides}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}, \tag{3.7}$$

$$d_{Minkowski}(x,y) = \sqrt[p]{\sum_{i=1}^{m}(x_i - y_i)^p}, \tag{3.8}$$

$$d_{Manhattan}(x, y) = \sum_{i=1}^{m} |x_i - y_i|, \tag{3.9}$$

gdzie m to liczba atrybutów warunkowych tabeli decyzyjnej, natomiast $x = [x_1, x_2, \ldots, x_m]$ i $y = [y_1, y_2, \ldots y_m]$ to wartości m atrybutów dwóch obiektów.

Przyjęta miara odległości powinna spełniać trzy podstawowe klasyczne warunki, zwane *aksjomatami metryki*.

- $d(x, y) = 0$, wtedy i tylko tedy, gdy $x = y$ (aksjomat tożsamości),
- $d(x, y) = d(y, x)$ (aksjomat symetrii),
- $d(x, y) \leq d(x, z) + d(z, y)$ (nierówność trójkąta).

Klasyczne miary, takie jak wymienione powyżej, wyznaczają odległość między obiektami na podstawie wartości atrybutów, którymi są zwykle wyniki pomiarów z czujników. Nie uwzględniają jednak zależności między obiektami, które zachodzą na wyższym poziomie abstrakcji, pomiędzy pojęciami, do których przynależą obiekty. W dalszej części rozprawy zaproponowano odległość opartą na ontologii pojęć, w której te zależności są uwzględniane (patrz Rozdz. 7).

3.3 Miary skuteczności klasyfikatorów

Istnieje szereg kryteriów, na podstawie których dokonywana jest ocena klasyfikatorów. Najczęściej rozważa się kryteria wymienione poniżej.

- Trafność klasyfikacji;
- Szybkość - ocenia się czas uczenia się oraz szybkość samego klasyfikowania;
- Skalowalność - ocenia się czy klasyfikatory mogą być tworzone i testowane na dużych zbiorach danych;
- Odporność (ang. *robustness*) na szum (ang. *noise*) czy też wartości brakujące w danych (ang. *missing values*);
- Zdolność wyjaśnienia podjętej decyzji;
- Złożoność modelu - określana na podstawie np. rozmiaru drzewa decyzyjnego.

Celem oceny jakości klasyfikatora w odniesieniu do badanych danych, w ogólnym przypadku tablica decyzyjna jest dzielona na dwie części (patrz np. [95]):

- *Tablica treningowa*, zwana także uczącą, zawierająca obiekty, na podstawie których algorytm uczy się klasyfikować obiekty do klas decyzyjnych;

- *Tablica testowa* służąca do oceny klasyfikatora utworzonego na podstawie części treningowej.

Takie podejście nazywane jest procedurą "trenuj i testuj" (ang. *train and test*). Zbiór treningowy i testowy powinny być reprezentatywne, tzn. np. rozkład występowania klas w obu zbiorach powinien odpowiadać rozkładowi występowania klas w zbiorze początkowym.

Najczęściej stosowaną ilościową metodą oceny klasyfikatorów przy jednokrotnym stosowaniu procedury "trenuj i testuj" jest metoda oparta na tzw. macierzy pomyłek, kontyngencji czy konfuzji (ang. *confusion matrix*). Macierz pomyłek jest tablicą dwuwymiarową, najczęściej kwadratową o wymiarach N_CxN_C, gdzie N_C jest liczbą klas i w polu (i, j), $i, j = 1, 2, .., N_C$ zawiera wartość oznaczającą liczbę przypadków z części testowej przynależnych do i-tej klasy decyzyjnej, które klasyfikator przyporządkował do klasy j-tej. Macierz ta jest podstawą do wyznaczenia wielu innych miar dokładności klasyfikacji.

Tablica 3.1 przedstawia macierz pomyłek dla przypadku dwóch klas decyzyjnych, np. przy klasyfikacji jakiegoś pojęcia. Komórki macierzy pomyłek zawierają następujące elementy (patrz [6]):

- TN (ang. *True Negatives*) - liczba prawidłowych klasyfikacji obiektów należących do przykładów negatywnych pojęcia w tabeli testowej;
- FP (ang. *False Positives*) - liczba nieprawidłowych klasyfikacji obiektów należących do przykładów negatywnych pojęcia w tabeli testowej;
- FN (ang. *False Negatives*) - liczba nieprawidłowych klasyfikacji obiektów należących do przykładów pozytywnych pojęcia w tabeli testowej;
- TP (ang. *True Positives*) - liczba prawidłowych klasyfikacji obiektów należących do przykładów pozytywnych pojęcia w tabeli testowej.

		Sklasyfikowane	
		Negatywne	Pozytywne
Rzeczywiste	Negatywne	TN	FP
	Pozytywne	FN	TP

Tablica 3.1: Macierz pomyłek.

Na podstawie danych z macierzy pomyłek dla dwóch klas decyzyjnych (negatywnej, oznaczonej 0 i pozytywnej zakodowanej za pomocą 1) skonstruowano kilka miar ułatwiających ocenę i porównywanie klasyfikatorów (patrz [6, 95, 118, 17]):

1. Dokładność ACC (ang. *accuracy*) - określająca odsetek przypadków sklasyfikowanych prawidłowo, wyznaczana za pomocą wzoru:

$$ACC = \frac{TN + TP}{TN + FN + FP + TP} \tag{3.10}$$

2. Czułość ACC_1 (ang. *accuracy for positive examples*), inaczej SN (ang. *sensitivity*) lub TPR (ang. *true positive rate*) - dana wzorem:

$$ACC_1 = SN = TPR = \frac{TP}{TP + FN} \tag{3.11}$$

3. Specyficzność ACC_0 (ang. *accuracy for negative examples*), inaczej SP (ang. *specificity*) lub TNR (ang. *true negative rate*) - dana wzorem:

$$ACC_0 = SP = FNR = \frac{TN}{TN + FP} \tag{3.12}$$

4. Pokrycie COV (ang. *coverage*), określa odsetek sklasyfikowanych obiektów ze wszystkich obiektów tablicy testowej (niektóre obiekty mogą nie zostać sklasyfikowane w ogóle) - dana wzorem:

$$COV = \frac{TN + FN + FP + TP}{\textit{liczba wszystkich obiektów}} \tag{3.13}$$

5. Pokrycie przykładów pozytywnych $PCOV$ (ang. *coverage for positive examples*), wyznaczane za pomocą wzoru:

$$PCOV = \frac{FN + TP}{\textit{liczba wszystkich przykładów pozytywnych}} \tag{3.14}$$

6. Pokrycie przykładów negatywnych $NCOV$ (ang. *coverage for negative examples*), wyznaczane za pomocą:

$$NCOV = \frac{TN + FP}{\textit{liczba wszystkich przykładów negatywnych}} \tag{3.15}$$

7. Precyzja przykładów pozytywnych PPV (ang. *positive predictive value*), inaczej dodatnia wartość predykcyjna, wyliczana ze wzoru:

$$PPV = \frac{TP}{TP + FP} \tag{3.16}$$

8. Precyzja przykładów negatywnych NPV (ang. *negative predictive value*), inaczej ujemna wartość predykcyjna, wyliczana ze wzoru:

$$NPV = \frac{TN}{TN + FN} \tag{3.17}$$

9. Współczynnik (odsetek) błędów ER (ang. *error rate*) – miara całkowitej liczby błędów popełnionych przez klasyfikator w odniesieniu do liczby wszystkich obiektów zadana jako:

$$ER = \frac{FN + FP}{TN + FN + FP + TP} \tag{3.18}$$

10. FPR (ang. *false positive rate*), czyli błąd I typu lub α:

$$FPR = \frac{FP}{FP + TN} = 1 - SP \tag{3.19}$$

11. FNR (ang. *false negative rate*), czyli błąd II typu lub β:

$$FNR = \frac{FN}{FN + TP} = 1 - SN \tag{3.20}$$

W literaturze dostępnych jest wiele innych metod oceny klasyfikatorów, takich jak krzywa ROC (ang. *Receiver Operating Characteristic curve*) (patrz [51, 149]), czy AUC (ang. *area under curve*). Charakterystyka ROC jest wykresem, który pokazuje zależność czułości SN od FPR podczas kalibrowania klasyfikatora. Te dwa współczynniki wyznaczane są na podstawie macierzy konfuzji i każdy binarny pojedynczy klasyfikator można przedstawić jako punkt w przestrzeni $(SN, 1-SP)$. Natomiast AUC jest współczynnikiem określającym powierzchnię pod krzywą ROC. Im większa powierzchnia, tym lepszy klasyfikator. Dla idealnego klasyfikatora wartość AUC wynosi 1

Jednokrotny podział losowy na dwa niezależne zbiory: uczący i testowy, celem oszacowania miar klasyfikacji stosuje się w przypadku dużych zbiorów danych, zawierających powyżej tysiąca obiektów [95]. Zwykle do zbioru testowego wybiera się losowo 20-30% obiektów z całego badanego zbioru danych. Dla danych o średnich rozmiarach (od 100 do kilku tysięcy obiektów) stosuje się zwykle technikę zwaną k-krotną oceną krzyżową lub kroswalidacją CV (ang. *k-fold cross-validation*). W metodzie tej zbiór danych jest losowo dzielony na k możliwie równych wzajemnie niezależnych części (najczęściej $k = 10$) i stosuje się $k-1$ podzbiorów jako części uczącej i pozostałej jako testującej. Sam klasyfikator konstruowany jest k-krotnie, a ocena klasyfikatora jest średnią wszystkich k ocen. Każda

część jest użyta $k-1$ razy do konstrukcji drzewa i 1 raz do testowania dokładności klasyfikacji. W przypadku małego zbioru danych wykorzystywana jest technika n-krotnej walidacji krzyżowej, zwana LOO (ang. *Leaving-One-Out*), w której liczba iteracji jest równa liczbie wszystkich obiektów n [65, 63]. Walidacja krzyżowa stanowi przykład próbkowania bez powtórzeń. Każdy obiekt jest testowany dokładnie jeden raz dla pewnego zbioru treningowego.

Estymacja miar może być bardziej wiarygodna, jeśli proces jest powtarzany dla różnych podzbiorów. Stosuje się w tym celu także metodę wielokrotnego repróbkowania (ang. *bootstrapping*), czyli losowanie przykładów ze zwracaniem z oryginalnego zbioru przykładów. Oryginalny zbiór jest próbkowany n razy tworząc zbiór treningowy o liczebności n. Ponieważ jest to losowanie ze zwracaniem, niektóre przykłady będą się powtarzać w zbiorze treningowym, a inne nie wystąpią (zbiór niewybranych elementów z języka angielskiego nazywa się zbiorem *out-of-bag*). Niewylosowane przykłady mogą tworzyć zbiór testowy, wykorzystywany do oceny dokładności klasyfikatora. Obiekt nie zostanie wybrany do zbioru treningowego z prawdopodobieństwem $1-\frac{1}{n}$. Prawdopodobieństwo tego, że pozostaje w zbiorze testowym wynosi:

$$\left(1-\frac{1}{n}\right)^2 \approx e^{-1} = 0.368 \tag{3.21}$$

Oznacza to, że zbiór treningowy zawiera ok. 63.2% przykładów. Z tego powodu mówi się o metodzie "0.632 bootstrap".

3.4 Metody selekcji cech

Jednym z problemów w zadaniach klasyfikacji jest wielowymiarowość obiektów przypisanych do poszczególnych klas. Wielowymiarowość stanowi poważne utrudnienie dla efektywności algorytmów eksploracji danych. Redukcja wymiarów może odbywać się poprzez proces selekcji cech (ang. *feature selection*), który polega na wybraniu możliwie dobrego podzbioru cech z pełnego zestawu wejściowego [63]. Jako "dobry" podzbiór cech uznaje się zestaw nie zawierający cech zbędnych. Zbędne cechy nie wprowadzają żadnej nowej informacji lub też nie mają żadnego związku z celem klasyfikacji, działają jak szum, powodują wydłużenie czasu uczenia, dlatego przed przystąpieniem do uczenia usiłuje się je wykryć i usunąć.

Wyróżnia się dwa rodzaje zbędnych cech: nieistotne (ang. *irrelevant*) i nadmiarowe (ang. *redundant*) [92]. Cechy nieistotne są cechami nieskorelowanymi z etykietami klas. Nie oznacza to, że zmienne są źle określone, pozbawione jakiejkolwiek wartości lub błędnie zmierzone. Mogą być po prostu niezwiązane z rozpatrywanym w danym momencie zagadnieniem. Cechy nadmiarowe natomiast to cechy, których wartości można wyliczyć z wartości pozostałych cech. Najprostszy przypadek nadmiarowości stanowi cecha będąca dokładnym powtórzeniem innej, tj.

dla każdego obiektu wartości tych dwóch cech są jednakowe. Odrzucenie zbędnych cech umożliwia zmniejszenie wymagań pamięciowych i złożoności czasowej algorytmów uczenia oraz poprawę zdolności uogólniania danego klasyfikatora, a więc polepszenie wyników klasyfikacji. Selekcja cech, poprzez wskazanie najistotniejszych atrybutów w zbiorze uczącym, prowadzi do skoncentrowania się algorytmu uczenia na najbardziej użytecznych aspektach danych.

Metody selekcji cech składają się zazwyczaj z czterech elementów, takich jak: generowanie podzbioru cech, walidacja podzbioru, kryterium zakończenia selekcji (stopu), ocena rezultatów [43]. Najczęstszym podejściem jest sekwencyjny przegląd zestawów cech według pewnej strategii i ocena jakości każdego zestawu. Strategia ta może polegać na przykład na dodawaniu jednej cechy wybranej losowo lub w szczególny sposób. Niestety przegląd wszystkich zestawów cech jest zwykle niemożliwy ze względu na czas selekcji rosnący wykładniczo z wymiarem danych. Dla m cech wejściowych istnieje $(2^m - 1)$ podzbiorów cech, a więc w praktyce pełny przegląd jest wykonalny jedynie dla zbioru nie posiadającego więcej niż kilkanaście cech. Dla takich zbiorów danych, pełny przegląd jest zalecany, jeżeli $n >> m$ (gdzie n jest licznością zbioru), inaczej takie podejście może doprowadzić do przeuczenia.

Generowanie podzbioru cech może odbywać się na różne sposoby. Podstawowymi strategiami są: przeszukiwanie w przód - strategia FSS (ang. *Forward Selection Strategy*) i przeszukiwanie wstecz - strategia BSS (ang. *Backward Selection Strategy, backward elimination*) [168]. W procedurze przeszukiwania w przód, w pierwszym kroku do pustego podzbioru atrybutów dodawana jest cecha uznana za najlepszą bez uwzględnienia zależności między cechami. W kolejnym kroku dodawany jest atrybut, który wraz z wybranym wcześniej tworzy najlepszą parę cech. Procedura ta przebiega iteracyjnie, aż do osiągnięcia kryterium zatrzymania. Ostateczną odpowiedzią jest zestaw najlepszy ze wszystkich rozpatrywanych. W przypadku zestawów równoważnych preferuje się ten, w którym ostatnio dołączona cecha, traktowana samodzielnie, oferuje mniejszy błąd klasyfikacji. Procedura przeszukiwania wstecz rozpoczyna się od pierwotnego zbioru cech. Następnie z podzbioru kolejno usuwane są cechy, w taki sposób, aby pomniejszony zestaw był w danym kroku najlepszy z możliwych. Rozszerzeniem strategii FSS i BSS jest strategia dwukierunkowa: w każdym kroku można albo dodawać, albo usuwać jedną cechę, w zależności od tego, co daje lepszy wynik.

Istnieją dwa różne podejścia do ewaluacji podzbiorów cech. Pierwsze określane jest jako metoda filtracyjna (ang. *filter*), ponieważ zbiór atrybutów jest filtrowany w celu utworzenia najbardziej obiecującego podzbioru przed rozpoczęciem eksploracji danych. Dla każdej cechy z osobna wyznaczany jest pewien współczynnik (indeks) określający jej jakość według przyjętego kryterium. Na podstawie wartości indeksów tworzone są rankingi cech. Istnieje wiele sposobów tworzenia indeksów, wśród których wyróżnia się metody oparte na korelacji wartości danej cechy

z etykietą klasy, odległościach pomiędzy ich rozkładami czy kryteriach stosowanych w drzewach decyzyjnych. Selekcja polega na wyborze najlepszych cech (pierwszych w rankingu) powyżej pewnego ustalonego progu, którym może być określona liczba cech, które należy pozostawić lub wartość indeksu oceniającego. Metody rankingowe z definicji nie uwzględniają zależności pomiędzy cechami, przez co mogą okazać się niewystarczające w przypadku występowania korelacji pomiędzy cechami. W drugim podejściu ocenia się poszczególne cechy z wykorzystaniem algorytmów uczenia maszynowego [168]. Stąd podejście to nazywa się metodą opakowującą (ang. *wrapper*), ponieważ algorytm uczenia zawiera się w procedurze selekcji cech. Rezultaty uzyskane z wykorzystaniem metod opakowujących zależą wyłącznie od jakości algorytmu uczącego i dopasowania algorytmu do określonego zadania klasyfikacyjnego. Ocena podzbiorów cech jest najczęściej dokonywana przy użyciu pewnego modelu klasyfikacyjnego, a miarą jakości podzbioru jest dokładność klasyfikatora, oszacowana przy użyciu walidacji krzyżowej.

W rozprawie zastosowano takie metody selekcji cech jak selekcja realizowana bezpośrednio przez drzewa decyzyjne oraz selekcja przez eksperta na podstawie wiedzy dziedzinowej w przypadku konstrukcji klasyfikatorów metodą k-NN. Metody indukcji drzew decyzyjnych są tak zaprojektowane, aby wybrać najlepszy atrybut podczas podziału każdego węzła i nie powinny - w teorii - wybierać atrybutów nieistotnych lub bezużytecznych. W praktyce jednak może być to trudne do osiągnięcia, gdyż z każdym podziałem maleje liczebność zbioru obiektów, na którym dokonywany jest wybór cech. Bardzo podatna na nieistotne cechy jest także metoda k najbliższych sąsiadów, ponieważ zawsze pracuje w lokalnym sąsiedztwie klasyfikowanego obiektu, biorąc pod uwagę zaledwie kilka przykładów uczących przy podejmowaniu każdej decyzji. Stąd w rozprawie zaproponowano selekcję cech opartą na wiedzy dziedzinowej.

3.5 Metody tworzenia klasyfikatorów dla pojęć czasowych

Właściwością procesów zachodzących w rzeczywistym świecie, poza ich złożonością, jest również ciągła zmienność w czasie. Zachodzą w nich nie tylko zmiany parametrów obiektów, ale mogą pojawiać się także nowe cechy. Eksploracja zbiorów uwzględniających czas stanowi dużo większe wyzwanie niż danych statycznych. Do takich danych należą zapisy Holtera, a więc 24-godzinne zapisy EKG, zawarte w głównych danych eksperymentalnych rozprawy. Podczas gdy analiza danych dotyczących pojedynczego punktu czasowego lub bez istotnego wpływu czasu na badane zjawisko sprowadza się przede wszystkim do określenia relacji między zbiorami obiektów, w danych czasowych pojawia się wiele innych zagadnień.

W modelowaniu złożonych rzeczywistych zjawisk i procesów mówi się o tzw. złożonych systemach dynamicznych CDS (ang. *Complex Dynamical Systems*) (patrz [10, 45, 20]). Stanowią one kolekcje złożonych obiektów charakteryzujących się ciągłymi zmianami parametrów w czasie oraz wzajemnymi oddziaływaniami między obiektami. Obiekty stanowiące CDS mogą ze sobą współpracować lub konkurować, bądź wykonywać mniej lub bardziej skomplikowane czynności. Przykładem takiego systemu może być pacjent w trakcie leczenia, ruch uliczny czy grupa robotów symulująca np. grę zespołową. Często opis zachowania takiego systemu CDS nie jest możliwy przy użyciu samych metod analitycznych, ponieważ obejmuje wiele rozmytych pojęć (patrz np. [78, 79, 123]). Pojęcia te dotyczą właściwości wybranych fragmentów systemu CDS i mogą być traktowane jako mniej lub bardziej złożone obiekty występujące w CDS. Celem wyciągania wniosków na temat globalnego stanu systemu CDS potrzebne są metody ekstrakcji takich fragmentów CDS. Stan CDS może być opisywany za pomocą informacji o przynależności złożonych obiektów wyodrębnionych z CDS do zdefiniowanych uprzednio złożonych pojęć, które opisują właściwości złożonych obiektów oraz relacje między obiektami. Ponadto, opis dynamiki systemu CDS wymaga obserwacji kolejnych zmian systemu w czasie tworząc historię jego zachowania, czyli sekwencję stanów systemu CDS obserwowanych w pewnym okresie czasu. Wynika stąd potrzeba rozwijania metod obserwacji zmian wybranych fragmentów systemu CDS oraz zmian relacji między nimi. W rozprawie do reprezentacji oraz obserwacji zmian złożonych obiektów występujących w systemie CDS stosowane są pojęcia czasowe. Pojęcia czasowe wyrażone są w języku naturalnym na dużo wyższym poziomie abstrakcji niż dane pochodzące z czujników. Przykładami takich pojęć są: zachowanie pacjenta w stanie zagrożenia życia czy bezpieczna jazda samochodem. Identyfikacja złożonych pojęć oraz ich zastosowanie do monitorowania stanu systemu CDS wymaga jednak wcześniejszej aproksymacji takich pojęć za pomocą klasyfikatorów na podstawie dostępnych danych sensorowych oraz wiedzy dziedzinowej.

Reprezentowanie czasu w danych

Identyfikacja złożonych pojęć odbywa się zwykle na podstawie pewnej reprezentacji wiedzy historycznej, używanej do przechowywania informacji na temat zmian wybranych parametrów i cech. Taka informacja jest zazwyczaj przedstawiana w postaci zbioru danych kolekcjonowanych podczas dłuższego czasu obserwacji złożonego systemu dynamicznego CDS (patrz [19, 21, 22, 111]).

Zbiory danych stosowane do przechowywania informacji na temat złożonych obiektów w systemie CDS mogą być reprezentowane za pomocą systemów informacyjnych SI. W takim podejściu złożone obiekty są reprezentowane przez wiersze (obiekty) systemu informacyjnego, a ich właściwości przez kolumny (atrybuty) systemu SI. Załóżmy dla potrzeb tej rozprawy, że obiekty ze zbioru U są opisane za pomocą skończonego zbioru atrybutów, reprezentujących cechy obiektów $A = \{a_1, a_2, ..., a_m\}$. Każdy atrybut $a \in A$ koresponduje z funkcją $a : U \to V_a$, zwaną funkcją oceny, gdzie V_a stanowi dziedzinę atrybutu a.

W rodzinie CDS wyróżnia się systemy jedno- i wieloobiektowe. Ponieważ w tym drugim przypadku różne elementy $u \in U$ mogą odnosić się do tego samego złożonego obiektu, dlatego wprowadza się identyfikatory pojedynczych złożonych obiektów. Ta informacja może być reprezentowana przez dodatkową kolumnę systemu informacyjnego oznaczoną przez a_{id}. Załóżmy, że wartości atrybutu a_{id} są uporządkowane liniowo. Zatem atrybut ten musi posiadać relację porządkującą zbiór wartości w porządku liniowym. Ponadto, wartości parametrów złożonych obiektów muszą być rejestrowane w różnych punktach czasowych. To z kolei wymusza zapisywanie, poza identyfikatorem obiektu, także identyfikatora punktu czasowego. Ta informacja może być zapisana w kolejnym dodatkowym atrybucie określanym jako a_t. Ponieważ zakładamy, że wartości atrybutu a_t są uporządkowane liniowo, więc również ten atrybut musi posiadać relację porządkującą liniowo zbiór wartości.

Standardowy system informacyjny SI przedstawiony w Rozdz. 2.3.1, wymaga zatem pewnych rozszerzeń. W tym celu definiuje się tzw. rozszerzony system informacyjny, zwany temporalnym systemem informacyjnym TIS (ang. *temporal information system*) (patrz [20, 150]).

Definicja 3.5.1 (Temporalny system informacyjny TIS)
Temporalny system informacyjny to 6-elementowa krotka:

$$\boldsymbol{TIS} = (U, A, a_{id}, \leq_{a_{id}}, a_t, \leq_{a_t}), gdzie:$$

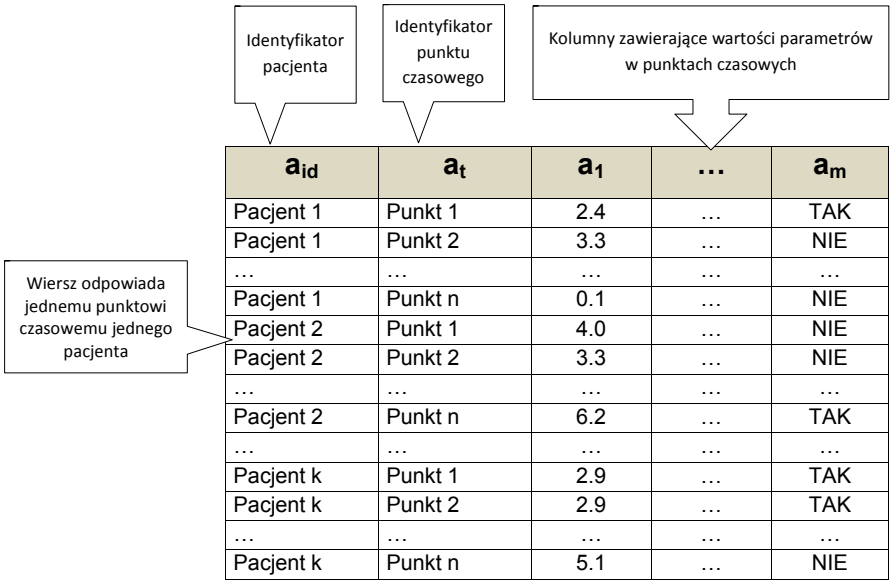
- *(U,A) to system informacyjny,*
- a_{id}, a_t *są wybranymi atrybutami ze zbioru A,*
- $\leq_{a_{id}}$ *jest relacją określającą liniowy porządek zbioru* $V_{a_{id}}$,
- $\leq_{a_t}$ *jest relacją określającą liniowy porządek zbioru* V_{a_t}.

Element $u \in U$ reprezentuje parametry złożonego obiektu o identyfikatorze $a_{id}(u)$ w punkcie czasowym $a_t(u)$. Obiekt $u_1 \in U$ poprzedza obiekt $u_2 \in U$ wtedy i tylko wtedy, gdy:

$$u_1 \neq u_2 \wedge a_{id}(u_1) = a_{id}(u_2) \wedge a_t(u_1) \leq_{at} a_t(u_2)$$

Przykładem temporalnego systemu informacyjnego jest system informacyjny, w którym obiekty reprezentują status pacjentów w różnych momentach obserwacji. Założenie o liniowości porządku nie jest obligatoryjne. Na przykład w [150] autor eksplorował sekwencje logów do stron www (temporalny system informacyjny). Aby pokazać, że z danej strony można przejść do kilku różnych, wprowadził relację częściowego porządku na atrybucie a_t.

Przykład 3.5.1 *Załóżmy, że mamy temporalny system informacyjny* $\boldsymbol{TIS} = (U, A, a_{id}, \leq_{aid}, a_t, \leq_{at})$*, którego obiekty reprezentują stany pacjentów w różnych punktach czasowych. Atrybuty ze zbioru A opisują parametry z czujników w danym punkcie, takie jak maksymalna częstotliwość pracy serca HR (ang. heart rate), liczba uniesień odcinka ST czy liczba tachykardii. Dana wartość atrybutu a_{id} stanowi jednoznaczny identyfikator danego pacjenta, natomiast atrybut a_t określa numer punktu czasowego, w którym dokonano rejestracji wartości parametrów (patrz Rys. 3.5).*



a_{id}	a_t	a_1	...	a_m
Pacjent 1	Punkt 1	2.4	...	TAK
Pacjent 1	Punkt 2	3.3	...	NIE
...	...	...	...	...
Pacjent 1	Punkt n	0.1	...	NIE
Pacjent 2	Punkt 1	4.0	...	NIE
Pacjent 2	Punkt 2	3.3	...	NIE
...	...	...	...	...
Pacjent 2	Punkt n	6.2	...	TAK
...	...	...	...	...
Pacjent k	Punkt 1	2.9	...	TAK
Pacjent k	Punkt 2	2.9	...	TAK
...	...	...	...	...
Pacjent k	Punkt n	5.1	...	NIE

Rysunek 3.5: Przykład temporalnego systemu informacyjnego TIS.

Pojęcia czasowe i ich aproksymacja

Problem przewidywania przynależności danego obiektu do złożonego pojęcia można traktować jako przykład problemu aproksymacji pojęć. Takie problemy mogą być modelowane za pomocą systemu złożonych obiektów i ich części oddziałujących wzajemnie na siebie. Systemy takie określa się jako złożone systemy dynamiczne CDS. Na przykład, w przypadku przewidywania odpowiedzi pacjenta na leczenie, pacjent może być traktowany jako badany złożony system dynamiczny, natomiast jego choroby jako złożone obiekty zmieniające się w czasie oraz wpływające na siebie. Pojęcia i metody ich aproksymacji stanowią użyteczne narzędzie do efektywnego monitorowania CDS. Każde pojęcie może być rozumiane jako sposób reprezentacji pewnych cech, właściwości złożonego obiektu.

Aproksymacja takich pojęć może odbywać się za pomocą parametrów (wartości sensorowych) zarejestrowanych dla pewnego zbioru złożonych obiektów. Jednak percepcja złożonych cech złożonych obiektów wymaga obserwacji takich obiektów przez dłuższy czas zwany oknem czasowym TW (ang. *time window*), gdzie okno czasowe może być rozumiane jako sekwencja obiektów danego temporalnego systemu informacyjnego dotycząca danego złożonego obiektu począwszy od określonego punktu czasowego przez określoną liczbę punktów czasowych. Niech $TW(\mathbf{TIS})$ oznacza rodzinę wszystkich okien czasowych systemu **TIS** oraz $card(W)$ oznacza długość okna czasowego $W \in TW(\mathbf{TIS})$. Rodzina wszystkich okien czasowych systemu **TIS** o długości równej s jest oznaczana jako $TW(\mathbf{TIS}, s)$. Elementy każdego okna czasowego $W \in TW(\mathbf{TIS}, s)$ są uporządkowane liniowo za pomocą relacji $\leq_{at}$, zatem każde okno czasowe może być traktowane jako uporządkowana sekwencja $W = (u_1, ..., u_s)$ obiektów ze zbioru U. Okno W może być opisywane formułą postaci: (i, b, s), gdzie $i \in V_{a_{id}}, b \in V_{a_t}$ oraz $s \in \mathbb{Z}_2$ dla $\mathbb{Z}_2$ będącego zbiorem liczb całkowitych większych lub równych 2. Dodatkowo każdy i-ty obiekt okna czasowego W oznacza się jako $W[i]$, gdzie $i \in \{1, ..., s\}$. Poniżej podano przykład ekstrakcji okna czasowego z temporalnego systemu informacyjnego.

Przykład 3.5.2
Rozważmy temporalny system informacyjny $\boldsymbol{TIS} = (U, A, a_{id}, \leq_{aid}, a_t, \leq_{at})$*, którego obiekty reprezentują stany pacjentów w różnych punktach czasowych. Atrybuty ze zbioru A opisują parametry z sensorów w danym punkcie. Niech obiekt (pacjent) o identyfikatorze 3 posiada 100 punktów czasowych o identyfikatorach od 1 do 100, przy założeniu że wartości atrybutu* a_t *są liczbami naturalnymi. Dla tego pacjenta można wyodrębnić okno czasowe określone formułą (3,51,20), która reprezentuje zachowanie obiektu od punktu czasowego oznaczonego identyfikatorem 51, aż do punktu czasowego oznaczonego 70.*

Do konstrukcji złożonych cech stosowane są *wzorce czasowe*. Przykładami takich wzorców może być: pierwsza w oknie wartość atrybutu a, pojawienie się w oknie

czasowym pewnej zadanej wartości czy wystąpienie kolejno po sobie w danym oknie określonych wartości dwóch parametrów a i b. Zatem każdy wzorzec czasowy jest zdeterminowany przez wartości pewnych sensorów. Zakłada się, że każdy wzorzec czasowy jest zdefiniowany przez eksperta na podstawie wiedzy dotyczącej danego złożonego systemu dynamicznego. Wzorce czasowe mogą być wykorzystane do zdefiniowania nowych cech, stosowanych do aproksymacji bardziej złożonych pojęć, zwanych *pojęciami czasowymi.*

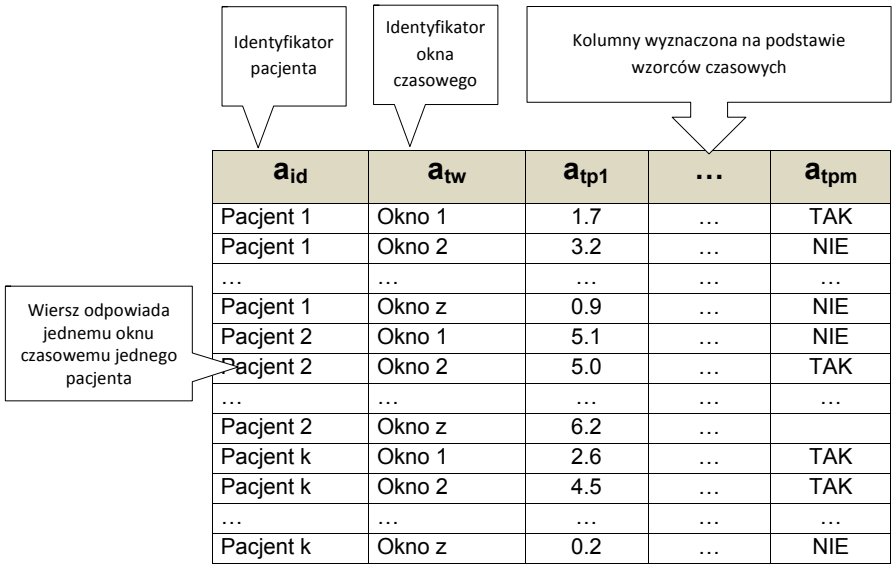
Zakładamy, że pojęcia te są wyszczególnione przez eksperta w danej dziedzinie. Intuicyjnie, każde pojęcie czasowe (zdefiniowane dla okna czasowego) opiera się na właściwościach obiektu obserwowanego w pewnych punktach czasowych. Z tego powodu pojęcia te mogą być aproksymowane za pomocą elementarnych pojęć opisujących cechy obiektów. Pojęcia czasowe zwykle są stosowane w pytaniach dotyczących stanu pewnych obiektów w danym oknie czasowym. Odpowiedzi na takie pytania są typu: *Tak, Nie* lub *Nie dotyczy.* Przykładowo, dla problemu leczenia pacjenta, można zdefiniować następujące złożone pojęcia: *Czy stan pacjenta ulega poprawie?, Czy pacjent reaguje pozytywnie na leczenie?* lub *Czy pacjent wymaga zmiany terapii?*

Zwykle problem aproksymacji pojęć jest formułowany jako problem uczenia indukcyjnego, tzn. problem poszukiwania przybliżonego opisu pojęcia C na podstawie skończonego zbioru przykładów $u \in U$, zwanego zbiorem uczącym. Aproksymacja powinna być możliwie jak "najbliżej" oryginalnego pojęcia, przy czym odległość może być określana dla różnych kryteriów, takich jak np.: dokładność czy długość opisu. Jeżeli dla danej tablicy decyzyjnej DT, $C \subset U$ jest pojęciem, które chcemy aproksymować, to atrybut decyzyjny d jest funkcją charakterystyczną pojęcia C. Zatem jeżeli $u \in C$, wówczas $d(u) = TAK$, w przeciwnym przypadku $d(u) = NIE$. Ogólnie, atrybut decyzyjny d może określać kilka rozłącznych pojęć. Wówczas, bez utraty ogólności zakłada się, że dziedzina decyzji d jest skończona i równa $V_d = \{1, 2, ..., n\}$. Dla dowolnego $k \in V_d$, zbiór $KLASA_k = \{u \in U : d(u) = k\}$ jest nazywana k-tą klasą decyzyjną w DT. Decyzja d wyznacza podział U na klasy decyzyjne, taki że $U = KLASA_1 \cup ... \cup KLASA_n$.

Przykład 3.5.3 *Przykładem problemu aproksymacji pojęć może być przewidywanie obecności zwężeń tętnic wieńcowych wymagających udrażniania u pacjentów z chorobą niedokrwienną serca na podstawie danych klinicznych oraz zapisu EKG metoda Holtera. Takie przewidywanie wymaga konstrukcji klasyfikatora, który na bazie dostępnej wiedzy przydziela pacjentów do zdefiniowanych klas decyzyjnych. Klasami decyzyjnymi w tym przypadku są: "Pacjenci bez istotnych zwężeń, niewymagający udrażniania" (klasa decyzyjna NIE) oraz "Pacjenci z istotnymi zwężeniami wymagający rewaskularyzacji" (klasa decyzyjna TAK). Klasyfikacja umożliwia więc podejmowanie decyzji dotyczących postępowania diagnostyczno-terapeutycznego w chorobie niedokrwiennej serca.*

Wzorce czasowe są często wykorzystywane w pytaniach zamkniętych z odpowiedziami: *Tak, Nie*. Przykładami takich wzorców czasowych dotyczących leczenia pacjenta mogą być: *Czy u pacjenta wystąpiło kiedykolwiek krwawienie z przewodu pokarmowego?, Czy przed utratą przytomności występowały zaburzenia rytmu serca?* lub *Czy doszło do przyspieszenia rytmu serca?*. Zakłada się, że wzorce czasowe powinny być zdefiniowane przez eksperta w danej dziedzinie.

Właściwości okien czasowych określone za pomocą wzorców czasowych mogą być reprezentowane w postaci specjalnego sytemu informacyjnego, zwanego *systemem informacyjnym okien czasowych* (patrz Rys. 3.6).



a_{id}	a_{tw}	a_{tp1}	...	a_{tpm}
Pacjent 1	Okno 1	1.7	...	TAK
Pacjent 1	Okno 2	3.2	...	NIE
...	...	...	...	...
Pacjent 1	Okno z	0.9	...	NIE
Pacjent 2	Okno 1	5.1	...	NIE
Pacjent 2	Okno 2	5.0	...	TAK
...	...	...	...	...
Pacjent 2	Okno z	6.2	...	
Pacjent k	Okno 1	2.6	...	TAK
Pacjent k	Okno 2	4.5	...	TAK
...	...	...	...	...
Pacjent k	Okno z	0.2	...	NIE

Rysunek 3.6: Schemat systemu informacyjnego dla okien czasowych.

Taka reprezentacja umożliwia zapisanie danych wszystkich obiektów, przykładowo pacjentów wraz z ich historią. Mogą w niej pojawić się także atrybuty statyczne, takie jak np. w przypadku pacjentów płeć, obecność przewlekłych chorób współistniejących, których wartości pozostają niezmienne w oknie czasowym.

Rozdział 4

Metoda I: Definiowanie cech w oparciu o wiedzę dziedzinową

Zawartość

Pierwsza metoda stanowi propozycję definiowania cech w oparciu o wiedzę dziedzinową za pomocą języka opartego na elementach logiki decyzyjnej i temporalnej. Cechy te zostaną wykorzystane tutaj do aproksymacji pojęcia, którym jest obecność istotnych zwężeń tętnic wieńcowych wymagających udrażniania u pacjentów z chorobą niedokrwienną serca na podstawie danych klinicznych oraz zapisu EKG metodą Holtera. Klasami decyzyjnymi w tym przypadku są: *Pacjenci bez istotnych zwężeń, niewymagający udrażniania* (klasa decyzyjna: NIE) oraz *Pacjenci z istotnymi zwężeniami wymagający rewaskularyzacji* (klasa decyzyjna: TAK).

Dane z 24-godzinnego zapisu Holtera uwzględniają upływ czasu. Pacjent jest charakteryzowany za pomocą wartości czujników w kolejnych punktach czasowych. Dłuższy okres czasu, w którym odbywa się rejestracja pomiarów w punktach czasowych stanowi okno czasowe. Dla wykorzystanych danych medycznych punkt czasowy dotyczył jednej godziny zapisu, natomiast okno obejmowało 24 godziny.

4.1 Definiowanie cech

Do definiowania cech odpowiednich dla okien czasowych wykorzystano eksperta, który w oparciu o wiedzę dziedzinową proponuje nie tylko same cechy, ale także sposób wyznaczania ich wartości w poszczególnych oknach czasowych. W rozprawie

proponuje się dwa rodzaje cech definiowanych przez eksperta dla okna czasowego W o długości s:

1. Cechy o wartościach liczbowych TP_W, wyznaczane jako:
   (a) Wartość funkcji agregującej dane w oknie czasowym: minimalna spośród wartości w oknie, maksymalna, średnia, odchylenie standardowe, pierwsza w oknie wartość, ostatnia w oknie wartość, wyznaczane kolejno za pomocą formuł:
   - $Min(W) = min(a(u))$ dla $u \in W$,
   - $Max(W) = max(a(u))$ dla $u \in W$,
   - $Mean(W) = \frac{\sum_{i=1}^{s} a(W[i])}{s}$,
   - $StdDev(W) = \sqrt{\frac{\sum_{i=1}^{s} [a(W[i]) - Mean(W)]^2}{s}}$,
   - $First(W) = a(W[1])$,
   - $Last(W) = a(W[s])$.

   (b) Dowolne wyrażenie arytmetyczne zgodne z podstawami arytmetyki operujące na wartościach z poprzedniego punktu.

2. Cechy o wartościach logicznych TP_B, wyznaczane jako:
   (a) Formuły relacyjne $<, \leq, >, \geq, =, \neq$ z udziałem cech o wartościach liczbowych, na przykład: $Max(W) > p$, gdzie p jest parametrem, którego wartość jest określana przez eksperta dziedzinowego.
   (b) Wyrażenia logiczne z udziałem kwantyfikatorów:
   - '`dla każdego`' $\forall$ (kwantyfikator ogólny). Cecha tego typu odpowiada na pytanie takie jak "czy dla każdego $u \in W$ prawdą jest, że $a(u) < p$?", gdzie $a \in A$ jest atrybutem, a p jest parametrem, którego wartość może być określona przez eksperta dziedzinowego lub zdefiniowana przez wartości innych formuł o wartościach numerycznych.
   - '`istnieje taki`' $\exists$ (kwantyfikator szczegółowy). Taka cecha odpowiada na pytanie typu "czy istnieje $u \in W$ takie, że $a(u) < p$?". Przykład: $\exists a(u) < First(W)$.

   (c) Wyrażenia logiczne z udziałem spójników: alternatywy, koniunkcji, negacji, implikacji $(\vee, \wedge, \sim, \Rightarrow)$. Przykładem tego typu cechy może być wyrażenie $(\forall a(u) = p_1 \wedge \exists b(u) > p_2)$, które przyjmuje wartość '*prawda*', gdy wartość atrybutu a dla każdego $u \in W$ wynosi p_1 i jednocześnie istnieje $u \in W$, dla którego wartość atrybutu b przekracza wartość p_2.

Tak zdefiniowane cechy nazywane będą wzorcami czasowymi TP (ang. *temporal patterns*) i mogą służyć do aproksymacji pojęć czasowych (patrz Rozdział 3.5). Funkcja `wzorzec(okno czasowe) = wartość wzorca`, która przypisuje wartość wzorca do okna czasowego opisuje stan lub zmiany stanu obiektu w punktach czasowych. Oto kilka przykładów wzorców czasowych.

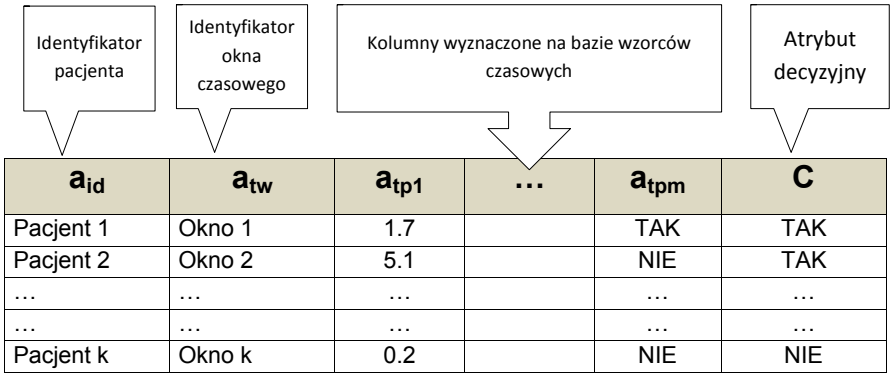
Przykład 4.1.1 *Przykłady cech dla okien czasowych.*

- *Dla atrybutu a zawierającego informację na temat częstotliwości rytmu serca pacjenta oraz wartości 70 i relacji $' >'$, formuła: $\exists a(u) > 70$ opisuje okna czasowe, w których wystąpił (chociaż raz) przyspieszony rytm serca (powyżej 70 uderzeń na minutę).*
- *Dla atrybutu a zawierającego informację na temat temperatury pacjenta oraz wartości 37 i relacji $' \leq'$, formuła: $\forall a(u) \leq 37$ dotyczy okien czasowych, w których temperatura ciała pacjenta nie przekraczała nigdy 37 stopni Celsjusza.*
- *Dla atrybutu a zawierającego informację na temat częstotliwości rytmu serca pacjenta, atrybutu b zawierającego liczbę zaburzeń rytmu (arytmii) i relacji $' >'$, formuła: $(\forall a(u) > 70) \wedge (\exists b(u) > 0)$ opisuje okna, w których u pacjenta cały czas występuje przyspieszony rytm serca oraz co najmniej 1 raz w oknie pojawia się arytmia.*

Dla eksperymentalnych danych medycznych jako wzorce czasowe zastosowano formuły wyznaczane na podstawie zapisu EKG metodą Holtera, takie jak np: pierwsza w oknie wartość średniego odstępu QT, średni w oknie maksymalny godzinowy rytm serca i odchylenie standardowe liczby zespołów QRS w oknie czasowym.

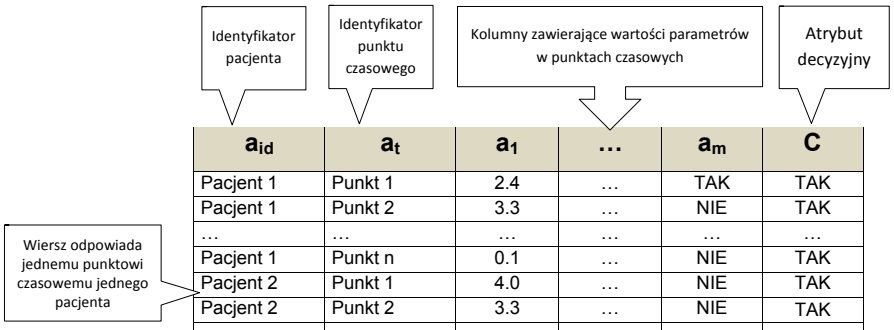
Cechy zdefiniowane powyżej stanowią jedynie przykład sposobu definiowania cech okien czasowych. Oczywiście istnieje możliwość definiowania innych cech tego typu, zawierających na przykład formuły weryfikujące, czy jakaś część punktów czasowych w oknie spełnia dany warunek (mniejszość lub większość).

Wzorce czasowe mogą być traktowane jako nowe cechy wykorzystywane do aproksymacji złożonych pojęć czasowych za pomocą klasyfikatorów. Celem zatem zastosowania klasyfikatorów do aproksymacji takich pojęć wymagana jest odpowiednia tablica decyzyjna, zwana *tablicą wzorców czasowych TPT* (ang. *temporal pattern table*), zawierająca atrybuty warunkowe wyznaczone na podstawie wzorców czasowych [15]. Każdy wiersz odpowiada jednemu oknu czasowemu obiektu (patrz Rys. 4.1). Należy zaznaczyć, że w ogólnym przypadku TPT dla jednego obiektu złożonego ma tyle rekordów, ile w DT było okien czasowych. Tablica wzorców czasowych jest konstruowana na bazie tablicy decyzyjnej DT składającej się z zarejestrowanych informacji na temat obiektów w złożonym systemie dynamicznym. Każdy wiersz tabeli DT zawiera informacje na temat parametrów pojedynczego obiektu w punkcie czasowym (patrz Rys. 4.2).



a_{id}	a_{tw}	a_{tp1}	...	a_{tpm}	C
Pacjent 1	Okno 1	1.7		TAK	TAK
Pacjent 2	Okno 2	5.1		NIE	TAK
...	...	...		...	...
...	...	...		...	...
Pacjent k	Okno k	0.2		NIE	NIE

Rysunek 4.1: Schemat tablicy wzorców czasowych TPT.



a_{id}	a_t	a_1	...	a_m	C
Pacjent 1	Punkt 1	2.4	...	TAK	TAK
Pacjent 1	Punkt 2	3.3	...	NIE	TAK
...	...	...	...	...	...
Pacjent 1	Punkt n	0.1	...	NIE	TAK
Pacjent 2	Punkt 1	4.0	...	NIE	TAK
Pacjent 2	Punkt 2	3.3	...	NIE	TAK
...	...	...	...	...	...
Pacjent 2	Punkt n	6.2	...	TAK	TAK
...	...	...	...	...	...
Pacjent k	Punkt 1	2.9	...	TAK	NIE
Pacjent k	Punkt 2	2.9	...	TAK	NIE
...	...	...	...	...	...
Pacjent k	Punkt n	5.1	...	NIE	NIE

Rysunek 4.2: Schemat tablicy DT.

Taka tablica może być traktowana jako zbiór danych zebranych na podstawie obserwacji zachowania złożonego systemu dynamicznego. Celem aproksymacji pojęcia czasowego C na podstawie tabeli (zbioru danych) DT należy skonstruować tabelę wzorców czasowych TPT następująco:

- Skonstruuj tablicę TPT z obiektami tablicy DT, dla których określono okna czasowe, tak aby liczba rekordów tablicy TPT odpowiadała liczbie okien czasowych;
- Każdy atrybut warunkowy tablicy TPT jest wyznaczany za pomocą wzorców czasowych zdefiniowanych przez eksperta do aproksymacji pojęcia C;
- Wartości atrybutu decyzyjnego (funkcji charakterystycznej pojęcia C) są proponowane przez eksperta.

Zakładamy, że każdy wzorzec czasowy jest wyznaczany na podstawie formuły zdefiniowanej przez eksperta w danej dziedzinie. Jest to zatem przykład zastosowania wiedzy dziedzinowej do poprawy jakości tworzonego klasyfikatora.

Następnie konstruowany jest klasyfikator dla tablicy TPT, który może aproksymować pojęcie czasowe C. Najpopularniejszą metodą konstrukcji klasyfikatorów jest indukcja reguł z przykładów (ang. *learning rule from examples*) (patrz [107]). Jednak reguły decyzyjne skonstruowane w ten sposób często są nieodpowiednie do klasyfikacji nowych, nieznanych wcześniej obiektów. Przykładowo, w przypadku tablicy decyzyjnej zawierającej atrybuty o wartościach ciągłych, szansa na rozpoznanie nowego obiektu przez reguły wygenerowane na podstawie tej tablicy jest mała, ponieważ wektor wartości atrybutów dla nowego obiektu może nie pasować do wygenerowanych reguł. Z tego powodu indukcja reguł powinna zostać poprzedzona dyskretyzacją wartości ciągłych. Ten problem jest intensywnie badany w pracy Hung S. Nguyena [101], z której zaczerpnięto metody dyskretyzacji rozważane w niniejszej rozprawie. Metody te oparte są na technikach zbiorów przybliżonych oraz wnioskowaniu boolowskim.

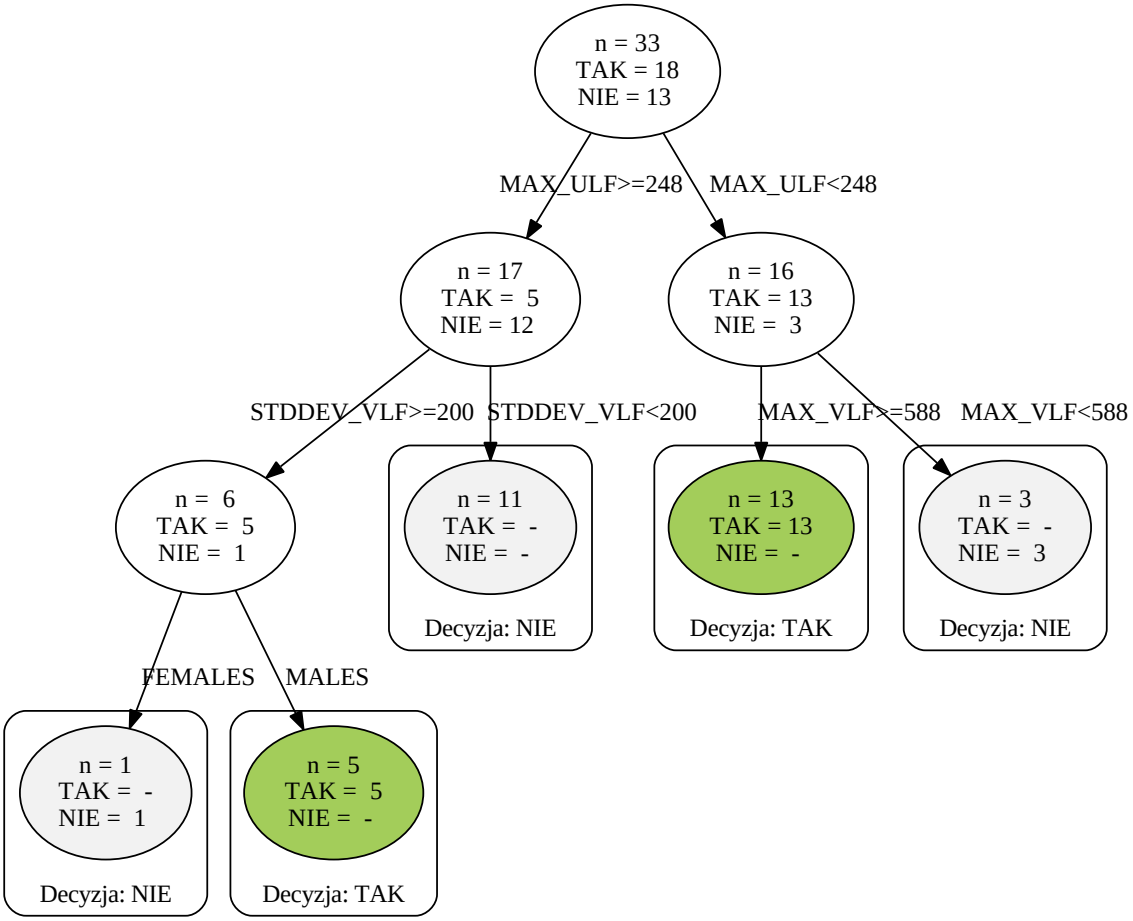
Sposób wyboru atrybutu i jego wartości wykorzystywanych do podziału zbioru obiektów stanowi kluczowy element rozważanej metody lokalnej dyskretyzacji i powinien się wiązać z analizą wartości atrybutu decyzyjnego w zbiorze trenującym (patrz Rozdział 3.1.1.) Celem zbudowania drzewa decyzyjnego poszukiwane jest najlepsze cięcie, najlepsze w sensie przyjętej miary.

4.2 Konstrukcja drzewa decyzyjnego z cechami zaproponowanymi przez eksperta

W proponowanej metodzie do budowy klasyfikatora dla danych medycznych wykorzystano miarę jakości cięć opisaną wzorem 3.1 (Rozdział 3.1.2), wyznaczaną jako liczba par obiektów, które są rozróżniane przez cięcie i należą do różnych klas decyzyjnych. Obliczając wartość tej miary dla wszystkich potencjalnych par *(atrybut, wartość)* można zachłannie wyznaczyć parę o najwyższej wartości miary i podzielić zbiór danych na dwie części na tej podstawie. Takie podejście określane będzie w rozprawie jako klasyczna metoda aproksymacji pojęć czasowych, a klasyfikator zbudowany za pomocą tej metody nazywany będzie dalej *CTree-Disc*. Jakość cięć może być wyznaczona dla dowolnego podzbioru danego zbioru obiektów.

Przykładowe tego typu drzewo otrzymane dla głównych danych tej rozprawy związanych ze stabilną chorobą wieńcową przedstawia Rys. 4.3. W każdym węźle drzewa podano liczbę obiektów tworzących węzeł oraz rozkład klas decyzyjnych.

Powyższe drzewo decyzyjne może być traktowane bezpośrednio jako klasyfikator, ponieważ obiekty testowe mogą być klasyfikowane poprzez określenie, do



Rysunek 4.3: Drzewo decyzyjne otrzymane metodą I do predykcji stenoz w CNS.

którego liścia drzewa przynależą. Jest to możliwe, ponieważ dzięki wyznaczonym podziałom węzłów, można prześledzić przynależność obiektu do ścieżki prowadzącej od korzenia do liścia, a następnie sklasyfikować obiekt do klasy decyzyjnej, której obiekty dominują w liściu.

Na przykład, dla drzewa z Rys. 4.3 w przypadku pacjenta z maksymalną w oknie czasowym wartością ULF, czyli pasma ultra niskiej częstotliwości widma HRV równej 112 ms^2 i maksymalną wartością VLF (pasmo bardzo niskiej częstotliwości) równą 256 ms^2, kierujemy się najpierw od korzenia drzewa do prawego poddrzewa, ponieważ pacjent pasuje do wzorca: $MAX\_ULF < 248$. W kolejnym etapie przechodzimy ponownie do prawego poddrzewa ($MAX\_VLF < 588$), który składa się z jednego węzła, zwanego liściem, gdzie kończy się ścieżka. Dopasowana do pacjenta ścieżka wskazuje, że tętnice wieńcowe tego pacjenta nie są istotnie zwężone przez miażdżycę. Natomiast dla mężczyzny o maksymalnej w oknie wartości

ULF równej 605 ms^2 i odchyleniu standardowym VLF równym 509.6, przewidujemy obecność miażdżycowego zwężenia tętnic wieńcowych.

Wzorce występujące w przedstawionym drzewie mają znaczenie kliniczne. Wygenerowane cięcia dotyczą płci oraz parametrów częstotliwościowych dobowej zmienności rytmu serca HRV (ang. *heart rate variability*), które są wykorzystywane w kardiologii do oceny ryzyka wystąpienia komorowych zaburzeń rytmu, a tym samym nagłego zgonu sercowego i prognozowania skuteczności leków antyarytmicznych. Ograniczona zmienność rytmu serca wiąże się ze zwiększonym ryzykiem zgonu. Natomiast wzrost mocy VLF jest zwiastunem zaburzeń rytmu [26, 153].

Innowacyjność proponowanej metody polega między innymi na łatwym w użyciu mechanizmie dodawania złożonej wiedzy dziedzinowej na potrzeby ekstrakcji cech.

Rozdział 5

Metoda II: Modyfikacja oceny jakości podziału w drzewie na podstawie macierzy odległości pomiędzy wartościami decyzji

Zawartość

W opisywanym podejściu *CTree-Disc* jako miarę jakości cięcia zastosowano liczbę par obiektów rozróżnianych przez cięcie, należących do rozdzielnych klas decyzyjnych. Jednak dodatkowa wiedza dotycząca charakterystyki pojęć reprezentowanych przez klasy decyzyjne może zostać użyta do lepszego rozróżniania par obiektów z różnych klas, prowadząc tym samym do poprawy jakości klasyfikacji. Przynależność do pojęcia zdefiniowanego jako "obecność zwężeń wymagających rewaskularyzacji" jest wyznaczana na podstawie wyników angiografii tętnic wieńcowych, czyli koronarografii. W oparciu o obraz koronarograficzny można wyróżnić chorobę jednonaczyniową, dwunaczyniową lub trójnaczyniową, w zależności od liczby zajętych tętnic. Taki podział anatomiczny choroby niedokrwiennej na 1-, 2- lub 3-naczyniową dostarcza użytecznych informacji prognostycznych i jest wykorzystywany w selekcji pacjentów planowanych do zabiegu udrażniania. Im więcej naczyń jest zmienionych, tym cięższy jest stan pacjenta i większe zagrożenie życia. Choroba trójnaczyniowa ma gorsze rokowania niż dwunaczyniowa, a ta z kolei zwykle gorsze niż jednonaczyniowa. Rozróżnienie liczby zmienionych naczyń wpływa również na postępowanie terapeutyczne. Ogólnie, pacjenci ze zwężeniem 1 lub 2 naczyń mogą uzyskać korzyści z zabiegu przezskórnej interwencji wieńcowej PCI

(ang. *pericutaneus coronary intervension*), podczas gdy choroba trójnaczyniowa wymaga zwykle zabiegu chirurgicznego na otwartym sercu - tzw. pomostowania CABG (ang. *coronary artery bypass graft*). W przypadku przewidywania obecności zwężeń wieńcowych, pacjenci bez istotnych zwężeń określani są jako nie należący do pojęcia, natomiast pacjenci z 1, 2, lub 3 zmienionymi naczyniami stanowią przykłady pozytywne pojęcia. A więc w klasycznym podejściu wykorzystuje się jedynie binarną informację o obecności istotnych zwężeń, oznaczając jako przykłady pozytywne tych pacjentów, u których stwierdzono jakiekolwiek istotne zwężenie tętnic.

Jednak takie kryterium może być zbyt proste, ponieważ grupa pacjentów z chorobą jednego lub więcej naczyń jest niejednorodna i zróżnicowana pod kątem objawów klinicznych. W szczególności z punktu widzenia badanych parametrów różnice mogą występować w zapisie EKG metodą Holtera. Kryterium to powoduje sprowadzenie wielowymiarowego problemu do jednego wymiaru. Ponadto w literaturze pojawiły się doniesienia, że EKG nie jest dobrym wskaźnikiem zajęcia tętnic wieńcowych z czułością na poziomie 51.5% [91] czy 62% [60] w prawidłowym wykrywaniu istotnych stenoz. Można spotkać doniesienia o niespójności zapisu EKG z obrazem angiograficznym (patrz np. [126, 91]). Zdarzają się także przypadki występowania całkowitego zamknięcia światła tętnicy wieńcowej, które nie są widoczne w zapisie EKG, np. w [56] u 16% pacjentów zapis EKG podczas całkowitej okluzji naczynia był prawidłowy.

W związku z wynikami klasycznego podejścia do predykcji stenozy (SN=78%, patrz Rozdz. 9.2) oraz doniesieniami na temat ograniczonych możliwości zapisu EKG w wykrywaniu zwężeń tętnic wieńcowych podjęto próbę zaimplementowania dodatkowej wiedzy dziedzinowej do modelu aproksymowanego pojęcia. Informacja na temat zróżnicowania wewnątrz klas decyzyjnych nie była dotychczas wykorzystywana. Taka wiedza może usprawnić poszukiwanie cięć lepiej rozróżniających pary obiektów z różnych klas decyzyjnych niż w metodzie *CTree-Disc*, prowadząc do uzyskania lepszych niż dla metody klasycznej wyników klasyfikacji. Jest to dowód na potrzebę szerszego zastosowania wiedzy dziedzinowej do formułowania, reprezentacji czy eksploracji pojęć.

5.1 Macierz wag do rozróżniania wewnętrznego zróżnicowania klas

Proponowane podejście polega na przypisywaniu wag cięciom, które rozróżniają obiekty z różnych klas decyzyjnych wykorzystując przy tym informacje na temat wewnętrznego zróżnicowania klas, np. dla problemu CNS na temat liczby zmienionych naczyń. Cięcie, które rozróżnia najmniej odległe stany, tj. pacjentów bez

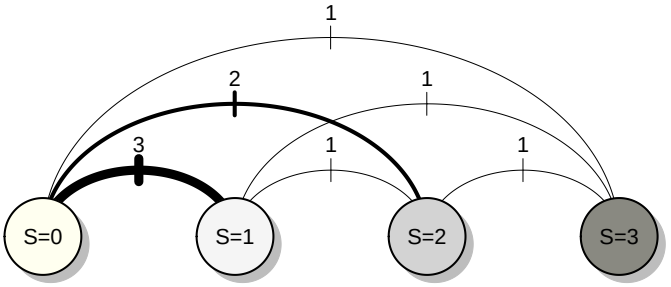
C		**NIE**	**TAK**		
	l.stenoz	**0**	**1**	**2**	**3**
NIE	**0**	0	3	2	1
TAK	**1**	3	0	1	1
	2	2	1	0	1
	3	1	1	1	0

Tablica 5.1: Wagi cięć uwzględniające liczbę zwężonych naczyń w CNS (reprezentacja tablicowa).

zmienionych naczyń od tych, którzy mają tylko jedno zwężenie, otrzymuje wagę o najwyższej wartości, ponieważ różnice między tymi grupami są najsubtelniejsze. Rozróżnienie przez cięcie pacjentów bez zmienionych naczyń od pacjentów z największą liczbą zwężeń, powoduje przypisanie takiemu cięciu najniższej wagi. Zakłada się, że wagi są określane przez eksperta w danej dziedzinie na podstawie odległości semantycznej pojęć. Jest to sposób na wyeksponowanie granicy pomiędzy obszarem negatywnym i pozytywnym złożonego pojęcia. Ponieważ jesteśmy zainteresowani rozróżnieniem przykładów negatywnych pojęcia od pozytywnych, dlatego największe wartości wag przydzielane są cięciom rozróżniającym pacjentów bez istotnych zwężeń tętnic, od tych, u których zwężenia występują.

Wykrycie subtelnych różnic między grupą pacjentów ze zmienionym jednym i dwoma naczyniami oraz dwoma i trzema nie wnosi użytecznej informacji do aproksymacji badanego pojęcia, więc wagi w tych przypadkach są najniższe. Dlatego zaproponowano wagi takie, jak w Tabeli 5.1, gdzie pierwszy wiersz i pierwsza kolumna oznaczają klasy decyzyjne, natomiast drugi wiersz i druga kolumna dotyczą liczby zwężonych tętnic wieńcowych. Graficzną prezentację wag poszczególnych cięć przedstawia także Rys. 5.1.

Wagi proponowane przez eksperta, są wykorzystywane do oceny jakości cięcia podczas rekurencyjnych podziałów zbiorów w węzłach drzewa lokalnej dyskretyzacji. Klasyfikator skonstruowany z wykorzystaniem wag oznaczany będzie jako klasyfikator *CTree-DiscW* . Łatwo zauważyć, że metoda wyznaczania miar jakości cięć w klasyfikatorze *CTree-Disc* (Metoda I) stanowi szczególny przypadek metody *CTree-DiscW*, w którym tablica wag 5.1 zawiera tylko dwie wartości: 0 lub 1. W klasyfikatorze *CTree-Disc* jakość cięcia jest liczbą par obiektów należących do przeciwnych klas decyzyjnych, wyznaczaną ze wzoru (3.1). Natomiast w metodzie *CTree-DiscW* miara ta wyznaczana jest jako suma wag wszystkich par obiektów należących do różnych klas z uwzględnieniem liczby zwężonych naczyń. Na przykład, jeżeli dane cięcie c dzieli zbiór obiektów na dwa podzbiory o liczebności M



Rysunek 5.1: Wagi cięć (reprezentacja graficzna). S - liczba stenoz

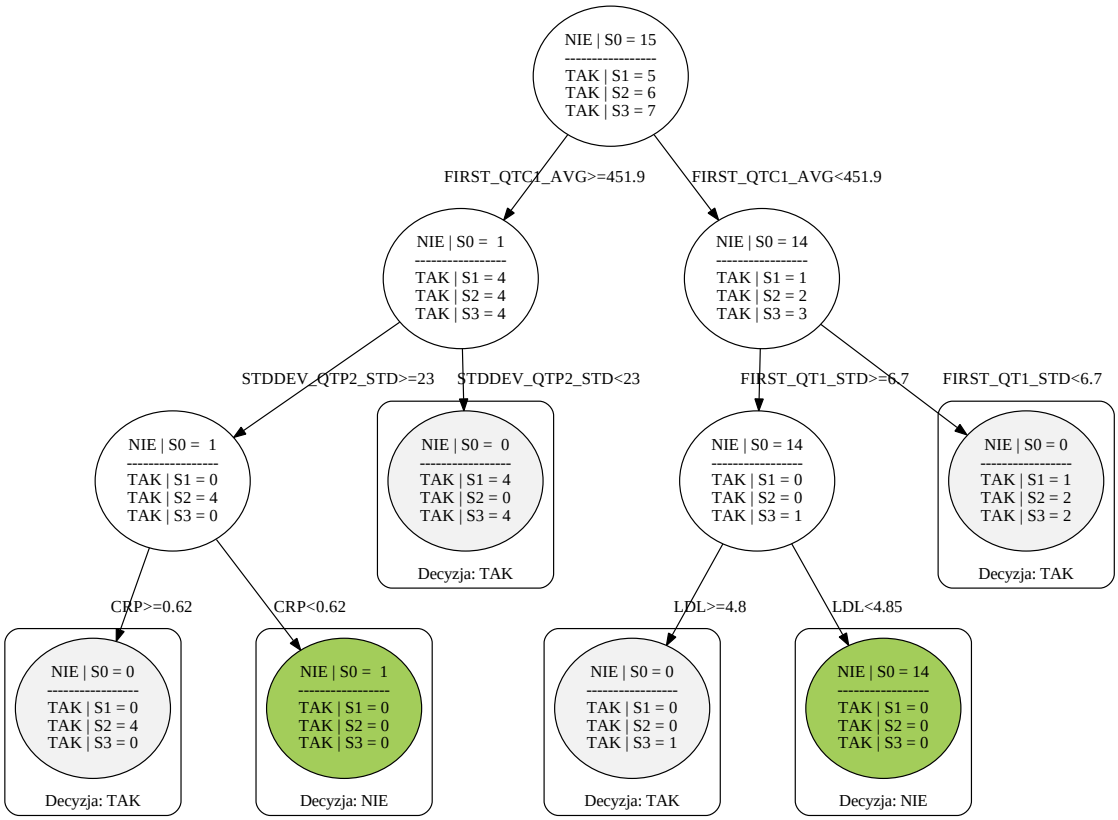
i N, a liczba obiektów bez zwężonych istotnie naczyń (klasa NIE) oraz ze stenozą 1, 2 lub 3 tętnic (klasa TAK) wynosi odpowiednio M_0, M_1, M_2, M_3 w jednej grupie oraz N_0, N_1, N_2, N_3 w drugiej, wówczas miara jakości cięcia jest wyznaczana następująco:

$$Q_{DiscW} = \sum_{i,j=0}^{3} w_{ij} M_i N_j \tag{5.1}$$

gdzie $w_{i,j}$ oznacza wagę cięcia rozróżniającego parę obiektów należących do różnych klas: bez stenozy oraz ze stenozą (1, 2 lub 3 istotnie zwężone naczynia).

Rysunek 5.2 przedstawia drzewo decyzyjne utworzone dla problemu przewidywania obecności zwężeń tętnic wieńcowych w medycznym zbiorze danych z wykorzystaniem klasyfikatora *CTree-DiscW*. W każdym węźle drzewa podano liczbę obiektów z poszczególnych klas decyzyjnych z uwzględnieniem informacji na temat wewnętrznego zróżnicowania klasy TAK, gdzie $S0$ oznacza liczbę obiektów bez zmienionych naczyń (liczba stenoz S równa 0), $S1$ liczbę obiektów z jedną stenozą, $S2$ z dwoma zwężeniami, a $S3$ liczbę obiektów węzła z trzema stenozami.

Przedstawione drzewo może być zastosowane do klasyfikacji obiektów świata rzeczywistego. Na przykład w przypadku pacjenta p, którego pierwszy w oknie czasowym średni czas trwania odstępu QTc w pierwszym odprowadzeniu EKG ($FIRST\_QTC1\_AVG$) wynosił 299 ms, odchylenie standardowe czasu trwania odstępu QT w pierwszym punkcie okna czasowego dla pierwszego odprowadzenia EKG ($FIRST\_QT1\_STD$) wynosiło 6.9, a poziom cholesterolu LDL 4.1 $mmol/l$, przemieszczamy się od korzenia, w dół do prawego poddrzewa, ponieważ pacjent pasuje do wzorca $FIRST\_QTC1\_AVG < 451$. W następnym kroku przechodzimy do lewego poddrzewa ze względu na dopasowanie do wzorca



Rysunek 5.2: Drzewo decyzyjne otrzymane metodą II do predykcji stenoz w CNS.

$FIRST\_QT1\_STD \geq 6.7$. Po odnalezieniu kolejnego cięcia, obiekt (pacjent) przemieszczany jest do prawego poddrzewa, który jest liściem. W liściu do obiektu p przypisywana jest ta klasa decyzyjna, do której przynależą wszystkie obiekty węzła. Ścieżka pasująca do nowego obiektu wskazuje, że w tętnicach tego pacjenta nie występują istotne zwężenia wymagające udrażniania.

Wzorce występujące w przedstawionym drzewie mają znaczenie kliniczne. Wygenerowane cięcia dotyczą głównie czasu trwania odstępu QT, który odzwierciedla czas trwania repolaryzacji komór mięśnia sercowego. Powszechnie stosuje się korekcję odstępu QT względem częstotliwości akcji serca celem zmniejszenia wpływu rytmu serca, wykonywaną za pomocą różnych wzorów (najczęściej używa się w tym celu wzoru Bazetta). Taki skorygowany odstęp QT nazywany jest odstępem QTc. Ocena odstępu QT i QTc pozwala rozpoznać groźną dla życia pacjenta patologię. Wydłużenie czy sporadycznie występujące skrócenie (poniżej 350 ms) odstępu QT może odpowiadać za zasłabnięcia, omdlenia lub nagły zgon sercowy NGS w wyniku częstoskurczu komorowego, migotania komór lub przedsionków [130]. Czas trwania QTc nie powinien przekraczać 450 ms u kobiet i 430 ms u mężczyzn [148]. Jedno

z cięć drzewa *CTree-DiscW* wynosi właśnie $451ms$ dla odstępu QTc, co wskazuje na zgodność uzyskanego wzorca z wiedzą dziedzinową.

Jednym z ograniczeń zaproponowanej metody jest konieczność posiadania dodatkowej informacji dotyczącej wewnętrznego zróżnicowania (rozwarstwienia) klas decyzyjnych. W zależności od wewnętrznej struktury tych klas, macierz wag może przyjmować różne rozmiary, co jest łatwo rozwiązywalne podczas implementacji algorytmu. Ostateczna jakość podziału obiektów na klasy decyzyjne w węźle drzewa jest wyznaczana na podstawie ich przynależności do wewnętrznych podklas. Nie ma jednak ograniczenia w stosowaniu tej metody tylko do atrybutów numerycznych, ponieważ modyfikacja jakości cięć może być zastosowana także do cech symbolicznych, w tym wyrażonych jako przedziały wartości.

Rozdział 6

Metoda III: Cięcia weryfikujące jako realizacja idei ekspertów dziedzinowych

Zawartość

W klasycznych metodach konstrukcji drzewa decyzyjnego dla każdego węzła wybierany jest zachłannie tylko jeden podział o najlepszej jakości względem ustalonej miary. Przy stosowaniu takiego podejścia dla przypadku danych z dużą liczbą atrybutów, pojawiają się istotne wątpliwości odnośnie zasadności takiego podejścia. Jego słabość leży w tym, że spośród być może wielu podziałów mających wysoką jakość, wybierany jest tylko jeden podział, a inne są pomijane. Jeśli liczba atrybutów jest mała i atrybuty niosą zróżnicowaną informację (np. w sensie zróżnicowanego obszaru pozytywnego względem atrybutu decyzyjnego), to takie podejście często jest skuteczne, tzn. prowadzi do wygenerowania efektywnych klasyfikatorów. Jednak w danych z dużą liczbą atrybutów może istnieć bardzo wiele atrybutów, które mają zbliżoną jakość, ale niosą znacząco różną informację o obiektach. Takie atrybuty będą tutaj nazywane *atrybutami nadmiarowymi (redundantnymi)*. Dobrze wiedzą o tym eksperci dziedzinowi (np. lekarze), którzy w swojej codziennej pracy zauważają taką nadmiarowość atrybutów i z niej korzystają, np. zwiększając pewność stawianych diagnoz poprzez użycie jednocześnie kilku atrybutów. Tymczasem wspomniana wyżej metoda zachłanna, spośród wielu atrybutów

redundantnych wybiera tylko jeden, pozostałe eliminując. W takim podejściu tracone są informacje zawarte w atrybutach, które są podobne pod względem jakości potencjalnych cięć, ale różnią się pod względem wiedzy dziedzinowej, którą reprezentują.

Jednak w praktyce, przy klasyfikacji obiektów testowych może pojawić się obiekt, który z jakichś powodów nie powinien być klasyfikowany zgodnie z podziałem wyznaczonym przez algorytm zachłanny. Np. może to być nietypowy obiekt z punktu widzenia wartości atrybutu wybranego w danym węźle drzewa przez algorytm lub w danych może pojawić się przekłamanie wartości atrybutu. Dlatego zdaniem ekspertów dziedzinowych możliwość klasyfikacji takiego obiektu przez wyznaczony zachłannie podział wymagałaby dodatkowego potwierdzenia za pomocą innych atrybutów, co w opisanej wyżej metodzie nie jest realizowane. Konsekwencją tego jest np. sytuacja, gdy dla danych mikromacierzowych liczących bardzo wiele atrybutów i niewiele obiektów metoda wyszukuje zaledwie kilka podziałów (na kilku atrybutach), które wystarczą do utworzenia drzewa pozwalającego na jednoznaczne sklasyfikowanie obiektów z próbki treningowej. Sytuacja taka jest bardzo trudna do zaakceptowania dla ekspertów dziedzinowych, którzy nie mogą się pogodzić z tak dużą redukcją wiedzy zawartej w atrybutach i niewykorzystywaniem redundantnych atrybutów podczas tworzenia drzewa.

Dlatego w rozprawie zaproponowano metodę weryfikacji podziałów w węźle drzewa przez inne podziały. Podstawowy pomysł polega na tym, aby na danym etapie wyszukiwania podziałów, po wyznaczeniu optymalnego podziału zbioru obiektów, wskazać dodatkowo rodzinę podziałów weryfikujących (wybierając do tego celu inne atrybuty niż atrybut użyty w optymalnym podziale), które możliwie podobnie jak podział optymalny oddzielają obiekty z różnych klas decyzyjnych.

Idea jaka temu przyświeca jest następująca. Jak wspomniano wyżej, optymalny podział jest narzędziem do cząstkowego sklasyfikowania obiektu testowego, tzn. dostarcza informacji, gdzie obiekt testowy powinien być posłany do sklasyfikowania: do prawego czy lewego poddrzewa. Każdy podział z rodziny podziałów weryfikujących dzieli podobnie obiekty z tablicy treningowej jak podział optymalny. Zatem jeśli pewien obiekt testowy zostanie skierowany przez podział optymalny do sklasyfikowania przez lewe drzewo, to istnieje domniemanie, że podobnie powinien go skierować także podział weryfikujący. Jeśli tak jest, to zwiększa się nasze przekonanie, że optymalny podział poprawnie sklasyfikował obiekt testowy. W przeciwnym przypadku, gdy np. podział optymalny kieruje obiekt testowy do lewego drzewa, a podział weryfikujący do prawego, to może to świadczyć o niepewności działania klasyfikatora i dlatego w takiej sytuacji zalecana jest ostrożność w zakresie planowania dalszego działania klasyfikatora. Ostrożność ta w przypadku proponowanej metody przejawia się w tym, że obiekt jest skierowany do klasyfikacji zarówno przez lewe jak i prawe poddrzewo. Po otrzymaniu wyników klasyfikacji rozstrzygany jest

ewentualny konflikt pomiędzy otrzymanymi decyzjami. Oczywiście podziałów weryfikujących może być więcej niż jeden i dlatego zarysowana wyżej metodologia musi to uwzględniać.

Osobnym zagadnieniem jest pytanie jak podziały weryfikujące ingerują w tworzenie samego drzewa dla tablicy treningowej. W klasycznej metodzie tworzenia drzewa (patrz Rozdział 3.1) na danym etapie tworzenia drzewa wyznaczany jest optymalny podział zbioru obiektów na dwa rozłączne zbiory, dla których w kolejnych etapach tworzone są osobne poddrzewa. Natomiast, w proponowanej metodzie podział zbioru obiektów może nie być rozłączny. Wprawdzie z jednej strony, tak jak poprzednio, optymalny podział dzieli zbiór obiektów treningowych na dwa rozłączne zbiory, ale może istnieć szereg obiektów, które pasują do wzorca zdefiniowanego na podstawie podziału optymalnego, ale nie pasują do któregoś ze wzorców zdefiniowanych dla podziałów weryfikujących. Podobnie mogą istnieć obiekty, które nie pasują do wzorca zdefiniowanego na podstawie podziału optymalnego, ale pasują do któregoś ze wzorców zdefiniowanych dla podziałów weryfikujących. O takich obiektach można powiedzieć, że już na etapie tworzenia drzewa jest wątpliwe, że wzorzec oparty na wyznaczonym podziale optymalnym jest odpowiedni do sklasyfikowania tego rodzaju obiektów. Dlatego podczas tworzenia drzewa obiekty takie zostaną dołączone zarówno do zbioru obiektów przeznaczonego do utworzenia lewego poddrzewa, jak i do zbioru obiektów przeznaczonego do utworzenia prawego poddrzewa. Odpowiada to intuicji, że uczenie się klasyfikowania tego rodzaju obiektów jest niejako odłożone w czasie i przekazane do obydwu poddrzew, gdzie dla ich sklasyfikowania zostaną policzone nowe podziały (być może lepiej dostosowane do tych obiektów niż optymalny podział policzony w bieżącym węźle drzewa).

W tym miejscu należy zauważyć, że opisany sposób wyboru podziałów weryfikujących w jednym węźle drzewa wpływa na wybór podziału optymalnego i podziałów weryfikujących w poddrzewie tego węzła. Jest to zatem inna sytuacja w porównaniu z możliwym podejściem alternatywnym, w którym najpierw utworzone zostałoby drzewo klasyczne (z użyciem tylko podziałów optymalnych), a następnie jego węzły zostały uzupełnione o podziały weryfikujące. W tym drugim przypadku samo drzewo byłoby takie samo jak klasyczne natomiast inaczej (ostrożniej) klasyfikowałoby obiekty testowe.

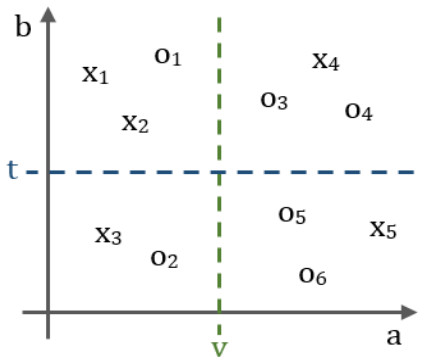
6.1 Wyznaczanie cięć weryfikujących

Dla przybliżenia idei weryfikacji podziałów w węźle drzewa przez inne podziały zostanie wyjaśnione i zilustrowane pojęcie jednoczesnego rozróżniania obiektów przez dwa cięcia. Para obiektów $(u_1, u_2) \in U \times U$ jest rozróżniana jednocześnie przez cięcie c_1 i c_2 definiujące wzorce $T1$ oraz $T2$ odpowiednio, jeżeli u_1 pasuje do

wzorca $T1$ oraz $T2$, natomiast u_2 nie pasuje ani do $T1$ ani do $T2$, lub odwrotnie, u_2 pasuje do wzorca $T1$ oraz $T2$, a u_1 nie. Na przykład, para (o_2, x_4) z Rys. 6.1 jest rozróżniana jednocześnie przez $c_1 = (a, v)$ oraz $c_2 = (b, t)$, natomiast pary (x_1, o_2) i (x_3, x_5) nie są rozróżniane jednocześnie przez cięcia c_1 oraz c_2.

Przez $Disc(c_1, c_2)$ oznaczana będzie liczba par obiektów z różnych klas decyzyjnych (dla danej tablicy decyzyjnej) rozróżnianych jednocześnie przez cięcia c_1 oraz c_2. Celem przedstawienia sposobu wyliczania wartości $Disc(c_1, c_2)$ rozważony zostanie zbiór obiektów z Rys. 6.1. Wzorzec $T(c_1)$ zdefiniowany przez cięcie $c_1 = (a, v)$ dzieli obiekty na podzbiory: $\{x_1, x_2, x_3, o_1, o_2\}$ oraz $\{x_4, x_5, o_3, o_4, o_5, o_6\}$. Natomiast wzorzec $T(c_2)$ zdefiniowany przez cięcie $c_2 = (b, t)$ dzieli zbiór obiektów na: $\{x_1, x_2, x_4, o_1, o_3, o_4\}$ oraz $\{x_3, x_5, o_2, o_5, o_6\}$. Celem wyliczenia wartości $Disc(c_1, c_2)$, należy sprawdzić, czy wzorce $T1$ i $T2$ są wzorcami prawymi czy lewymi. Zakładając, że $T(c_1) = TL(c_1)$ oraz $T(c_2) = TL(c_2)$, czyli obydwa wzorce są wzorcami lewymi, liczba obiektów pasujących jednocześnie do wzorca $T(c_1)$ i $T(c_2)$ wynosi 2 i są to obiekty: x_3, o_2 (jeden obiekt z klasy X i jeden z klasy O). Natomiast liczba obiektów niepasujących jednocześnie do wzorca $T(c_1)$ ani do $T(c_2)$ wynosi 3 i są to obiekty: x_4, o_3, o_4 (jeden obiekt z klasy X i dwa z klasy O). Zatem liczba par obiektów z różnych klas decyzyjnych rozróżnianych jednocześnie przez cięcie c_1 oraz c_2 jest dana wzorem: $Disc(c_1, c_2) = 1 \cdot 2 + 1 \cdot 1 = 3$.

W przypadku natomiast, gdy np.: $T(c_1) = TL(c_1)$, a $T(c_2) = TR(c_2)$ (pierwszy wzorzec jest lewy, a drugi prawy), liczba obiektów pasujących jednocześnie do wzorca $T(c_1)$ i $T(c_2)$ wynosi 3 i są to obiekty: x_1, x_2, o_1 (dwa obiekty z klasy X i jeden z klasy O). Z kolei liczba obiektów niepasujących jednocześnie do wzorca $T(c_1)$ ani $T(c_2)$ wynosi 3 i są to obiekty: x_5, o_5, o_6 (jeden obiekt z klasy X i dwa z klasy O). Zatem liczba par obiektów z różnych klas decyzyjnych rozróżnianych jednocześnie przez cięcie c_1 oraz c_2 wynosi teraz: $Disc(c_1, c_2) = 2 \cdot 2 + 1 \cdot 1 = 5$. Jest więc różna od poprzedniego przypadku, gdy obydwa wzorce były uznane za lewe.



Rysunek 6.1: Wizualizacja cięć w przestrzeni dwuwymiarowej.

Podczas konstrukcji drzew z cięciami weryfikującymi, dodatkowe cięcia są wyznaczane na podstawie wybranej miary jakości (porównaj Sekcja 3.1.2, s. 55) [50]. Algorytm 6.1.1 wyznacza cięcia weryfikujące używając trzech zaprezentowanych miar jakości dla wyznaczonego wcześniej cięcia optymalnego $p = (b, w)$, przy założeniu, że cięcia weryfikujące są wyznaczane na atrybutach numerycznych innych niż b. Dla ułatwienia rozważań załóżmy, że w danych są tylko dwie klasy decyzyjne C_0 i C_1. Podejście można oczywiście łatwo uogólnić na przypadek więcej niż dwóch klas decyzyjnych.

Algorytm 6.1.1: Wyznaczanie cięć weryfikujących na atrybutach numerycznych

WEJŚCIE: Tablica decyzyjna $\mathbf{A} = (U, A \cup \{d\})$ z klasami decyzyjnymi C_0 i C_1, cięcie $p = (b, w)$, parametr t lub t_w (zależnie od przyjętej miary; porównaj wzory 6.1-6.3)

WYJŚCIE: Kolekcja cięć weryfikujących dla cięcia p z wyselekcjonowaniem dla każdego cięcia weryfikującego $c = (a, v)$ właściwego prawego lub lewego wzorca $TL(c)$ lub $TR(c)$

begin

dla każdego *atrybutu* $a \in A$, *takiego że* $a \neq b$ **wykonaj**

1 Posortuj wartości atrybutu a, jeśli jest atrybutem o wartościach co najmniej ze skali porządkowej

2 Przeglądając wartości atrybutu a wyznacz dla każdego pojawiającego się cięcia c następujące liczby oraz umieść je w pamięci o cięciach M:

- $V_L(a, c, C_0)$ - liczba obiektów klasy decyzyjnej C_0 o wartościach atrybutu a mniejszych od v,
- $V_L(a, c, C_1)$ - liczba obiektów klasy decyzyjnej C_1 o wartościach atrybutu a mniejszych od v,
- $L(a, c, C_0)$ - liczba obiektów klasy decyzyjnej C_0 o wartościach atrybutu a mniejszych od v i jednocześnie pasujących do wzorca T_p
- $L(a, c, C_1)$ - liczba obiektów klasy decyzyjnej C_1 o wartościach atrybutu a mniejszych od v i jednocześnie pasujących do wzorca T_p.
- $V_H(a, c, C_0)$ - liczba obiektów klasy decyzyjnej C_0 o wartościach atrybutu a większych lub równych v,
- $V_H(a, c, C_1)$ - liczba obiektów klasy decyzyjnej C_1 o wartościach atrybutu a większych lub równych v,
- $H(a, c, C_0)$ - liczba obiektów klasy decyzyjnej C_0 o wartościach atrybutu a większych lub równych v i jednocześnie pasujących do wzorca $\neg T_p$,
- $H(a, c, C_1)$ - liczba obiektów klasy decyzyjnej C_1 o wartościach atrybutu a większych lub równych v i jednocześnie pasujących do wzorca $\neg T_p$.

3 Przeglądając pamięć M wyznacz jakości zapamiętanych cięć w sposób odpowiedni dla wybranej miary jakości:

Algorytm 6.1.1: Wyznaczanie cięć weryfikujących na atrybutach numerycznych - cd.

:

3 *DiscPairs*:
1. Wyznacz liczbę par obiektów rozróżnianych jednocześnie przez c i p:
$Disc(p,c) = max\{QL(c), QR(c)\}$, gdzie:
$QL(c) = L(a,c,C_0) \cdot H(a,c,C_1) + L(a,c,C_1) \cdot H(a,c,C_0)$,
$QR(c) = \big(V_L(a,c,C_0) - L(a,c,C_0)\big) \cdot \big(V_H(a,c,C_1) - H(a,c,C_1)\big)$
$+\big(V_L(a,c,C_1) - L(a,c,C_1)\big) \cdot \big(V_H(a,c,C_0) - H(a,c,C_0)\big)$;
oraz przypisz $T(c) = TL(c)$, jeżeli $QL(c) > QR(c)$, wpw
$T(c) = TR(c)$.
2. Wylicz jakość cięcia dla a na podstawie wzoru 6.1.
3. Wyznacz najlepsze cięcie, takie że wartość $QV_{Disc}(p,c)$
jest największa i większa od 0.
Entropia:
1. Wyznacz moc zbioru W: $|W| = min\{|W_L|, |W_R|\}$, gdzie:
$|W_L| = L(a,c,C_0) + L(a,c,C_1) + H(a,c,C_0) + H(a,c,C_1)$,
$|W_R| = |\mathbf{A}| - |W_L|$
oraz przypisz $T(c) = TL(c)$, jeżeli $|W_L| > |W_R|$, wpw $T(c) = TR(c)$.
2. Wyznacz następujące moce uniwersów podtablic tablicy $\mathbf{A}$
określonych przez cięcie c:
$|\mathbf{A}(T_c)| = V_L(a,c,C_0) + V_L(a,c,C_1)$,
$|\mathbf{A}(\neg T_c)| = V_H(a,c,C_0) + V_H(a,c,C_1)$.
3. Wylicz jakość cięcia c wg wzoru 6.2.
4. Wyznacz najlepsze cięcie według miary $QV_{Entropy}(p,c)$.
Gini:
1. Wyznacz moc zbioru W: $|W| = min\{|W_L|, |W_R|\}$, gdzie:
$|W_L| = L(a,c,C_0) + L(a,c,C_1) + H(a,c,C_0) + H(a,c,C_1)$,
$|W_R| = |\mathbf{A}| - |W_L|$
oraz przypisz $T(c) = TL(c)$, jeżeli $|W_L| > |W_R|$, wpw $T(c) = TR(c)$.
2. Wyznacz następujące moce uniwersów podtablic tablicy $\mathbf{A}$
określonych przez cięcie c:
$|\mathbf{A}(T_c)| = V_L(a,c,C_0) + V_L(a,c,C_1)$,
$|\mathbf{A}(\neg T_c)| = V_H(a,c,C_0) + V_H(a,c,C_1)$.
3. Wylicz jakość cięcia c wg wzoru 6.3.
4. Wyznacz najlepsze cięcie według miary $QV_{Gini}(p,c)$.

endfch

end

Opis algorytmu: W kroku pierwszym wartości atrybutu a, dla którego poszukiwane są cięcia weryfikujące, są sortowane. Drugi krok, dla kolejnych potencjalnych cięć $c = (a, v)$ przy ustalonym cięciu głównym $p = (b, w)$ na atrybucie b wyznacza liczebność następujących grup obiektów: $\{u \in U : a(u) < v \wedge dec(u) = C_0\}$, $\{u \in U : a(u) < v \wedge dec(u) = C_1\}$, $\{u \in U : a(u) < v \wedge (u \textit{ pasuje do } T_p) \wedge dec(u) = C_0\}$, $\{u \in U : a(u) < v \wedge (u \textit{ pasuje do } T_p) \wedge dec(u) = C_1\}$, $\{u \in U : a(u) \geq v \wedge dec(u) = C_0\}$, $\{u \in U : a(u) \geq v \wedge dec(u) = C_1\}$, $\{u \in U : a(u) \geq v \wedge (u \textit{ pasuje do } \neg T_p) \wedge dec(u) = C_0\}$, $\{u \in U : a(u) \geq v \wedge (u \textit{ pasuje do } \neg T_p) \wedge dec(u) = C_1\}$. Następnie, w kroku trzecim wyliczana jest jakość wszystkich potencjalnych cięć weryfikujących i na tej podstawie wybierane są najlepsze, spełniające warunek posiadania odpowiednio dobrej jakości, zależnie od ustawień parametrów. W przypadku miary *DiscPairs*, jakość cięcia weryfikującego musi wynosić co najmniej t procent jakości cięcia głównego. Jakość cięcia weryfikującego c jest definiowana jako liczba par obiektów z różnych klas decyzyjnych rozróżnianych jednocześnie przez cięcie główne i weryfikujące, a więc między następującymi dwiema grupami obiektów: między $\{u \in U : a(u) < v \wedge (u \textit{ pasuje do } T_p)\}$ a $\{u \in U : a(u) \geq v \wedge (u \textit{ pasuje do } \neg T_p)\}$ lub między $\{u \in U : a(u) < v \wedge (u \textit{ pasuje do } \neg T_p)\}$ a $\{u \in U : a(u) \geq v \wedge (u \textit{ pasuje do } T_p)\}$. Większa z tych dwóch wartości stanowi o jakości cięcia weryfikującego. Jeżeli większa liczba obiektów z różnych klas jest rozróżniania w pierwszej parze grup obiektów, wówczas do wzorca cięcia weryfikującego $T(c)$ przypisywany jest lewy wzorzec, inaczej wzorzec prawy. Dla pozostałych miar, wybierane są cięcia o najmniejszej różnicy ważonych sum entropii czy współczynników Giniego między zbiorami wyznaczonymi przez cięcia p oraz c. Odrzucane są natomiast cięcia, dla których liczebność zbioru W jest zbyt duża (jej odsetek w całym zbiorze A jest większy od t_w). Wielkość tego zbioru wskazuje bowiem na odmienny podział obiektów węzła przez obydwa cięcia: główne i weryfikujące.

Zakładając, że pamięć M dotycząca cięć i ich parametrów jest pamięcią o dostępie w czasie stałym, Algorytm 6.1.1 działa w czasie $O(m \cdot n \cdot log\ n)$ (ze względu na sortowanie obiektów względem wartości atrybutu a), gdzie m jest liczbą atrybutów, n jest liczbą obiektów.

Do ustalonego cięcia głównego może być też dobrane cięcie weryfikujące na atrybucie symbolicznym. Algorytm wyszukujący takie cięcie dla ustalonego atrybutu polegałby na przeglądaniu wszystkich wartości tego atrybutu i ustaleniu każdej z tych wartości jako cięcia weryfikującego. Dla każdego takiego cięcia należałoby przeglądać wartości atrybutu wszystkich obiektów celem wyznaczenia jakości cięcia. Złożoność takiego wyznaczania cięcia weryfikującego byłaby zależna jedynie od iloczynu liczby wartości tego atrybutu i liczby wszystkich obiektów.

6.2 Konstruowanie drzewa decyzyjnego z cięciami weryfikującymi

W rozdziale przedstawiono algorytm tworzenia drzewa decyzyjnego, który formalizuje powyższe rozważania (Algorytm 6.2.1). Ze względu na to, że algorytm ten wykorzystuje podziały weryfikujące, drzewo decyzyjne, które utworzy ten algorytm będzie nazywane *V-drzewem decyzyjnym* lub *V-drzewem*. Opisywane w pracy V-drzewa decyzyjne są dychotomiczne ze względu na każdy podział (optymalny i weryfikujące) stosowany przy ich budowie. Klasyfikator skonstruowany za pomocą *V-drzewa* będzie nazywany *VTree* klasyfikatorem.

W każdym węźle drzewa (który nie jest liściem), po wyznaczeniu optymalnego podziału zbioru obiektów wybierana jest rodzina podziałów podobnych do optymalnego, a przy tym wykorzystująca inne atrybuty. Oczywiście pojęcie podobieństwa zależy od miary, która jest zastosowana do określenia najlepszego podziału. W przypadku miary *DiscPairs* podobieństwo oznacza odróżnianie par obiektów z różnych klas decyzyjnych jak najbardziej zbliżonych do par odróżnianych przez optymalny podział [12, 11]. Natomiast w przypadku miar tworzonych na podstawie zysku informacji czy indeksu Giniego, podziały weryfikujące powinny dzielić zbiór obiektów w możliwie podobny sposób jak główny (optymalny) podział.

Na wejściu Algorytmu 6.2.1 należy podać tablicę decyzyjną $\mathbf{A} = (U, A \cup \{d\})$ oraz parametr k należący do liczb naturalnych oznaczający maksymalną liczbę podziałów weryfikujących oraz próg t definiujący minimalny wymóg stawiany każdemu podziałowi weryfikującemu. Wspomniany w powyższym algorytmie warunek stopu jest taki sam, jak w algorytmie omawianym w Rozdziale 3.1.3.

Do wyznaczania *V-drzewa* zastosowano trzy miary opisane w Rozdziale 3.1.2: miarę opartą na liczbie rozróżnianych par obiektów (*DiscPairs*), zysk informacji oraz indeks Giniego. W zależności od zastosowanej miary, optymalizowane są następujące kryteria:

- DiscPairs (kryterium jest maksymalizowane):

$$QV_{Disc}(p, p_i) = \begin{cases} 0 & \text{dla } \frac{Disc(p,p_i)}{Disc(p)} \leq t \\ \frac{Disc(p,p_i)}{Disc(p)} & \text{w przeciwnym wypadku} \end{cases} \tag{6.1}$$

  gdzie p jest cięciem optymalnym, $Disc(p)$ oznacza liczbę par obiektów z różnych klas decyzyjnych rozróżnianych przez cięcie p, natomiast $Disc(p, p_i)$ liczbę par obiektów z różnych klas decyzyjnych rozróżnianych jednocześnie przez cięcie p jak i p_i (dla $i = 1, ..., k$). W przeprowadzonych eksperymentach dotyczących omawianej metody wartość progu t ustawiono na 0.9.

- Miara oparta na entropii (kryterium jest minimalizowane):

$$QV_{Entropy}(p, p_i) = \begin{cases} 1 & \text{dla } \frac{|W|}{|\mathbf{A}|} \geq t_w \\ \big|ES\big(\mathbf{A}(T_p), \mathbf{A}(\neg T_p)\big) - & \\ \ ES\big(\mathbf{A}(T_{p_i}), \mathbf{A}(\neg T_{p_i})\big)\big| & \text{w przeciwnym wypadku} \end{cases} \tag{6.2}$$

  gdzie $|\ .\ |$ oznacza wartość bezwzględną oraz

  - W jest zbiorem obiektów, które nie pasują do wzorców T_p i T_{p_i} jednocześnie, jak również wzorców $\neg T_p$ oraz $\neg T_{p_i}$ (dla $i = 1, \ldots, k$),
  - t_w jest ustalonym progiem (t_w wynosił 0.1 oraz 0.05 w eksperymentach dla danych mikromacierzowych i pozostałych, odpowiednio),
  - $ES\big(\mathbf{A}(T_q), \mathbf{A}(\neg T_q)\big) = \frac{\big|\mathbf{A}(T_q)\big|}{|\mathbf{A}|} \cdot Entropia\big(\mathbf{A}(T_q)\big) + \frac{\big|\mathbf{A}(\neg T_q)\big|}{|\mathbf{A}|} \cdot Entropia\big(\mathbf{A}(\neg T_q)\big)$ jest ważoną sumą entropii cięć p i p_i, odpowiednio ($q \in \{p, p_1, \ldots, p_k\}$).

- Miara oparta na indeksie Giniego (kryterium jest minimalizowane):

$$QV_{Gini}(p, p_i) = \begin{cases} 1 & \text{dla } \frac{|W|}{|\mathbf{A}|} \geq t_w \\ \big|GS\big(\mathbf{A}(T_p), \mathbf{A}(\neg T_p)\big) - & \\ \ GS\big(\mathbf{A}(T_{p_i}), \mathbf{A}(\neg T_{p_i})\big)\big| & \text{w przeciwnym wypadku} \end{cases} \tag{6.3}$$

  gdzie:

  - W jest zbiorem obiektów, które nie pasują do wzorców T_p i T_{p_i} jednocześnie, jak również wzorców $\neg T_p$ oraz $\neg T_{p_i}$ (dla $i = 1, \ldots, k$),
  - t_w jest ustalonym progiem (t_w wynosił 0.1 oraz 0.05 w eksperymentach dla danych mikromacierzowych i pozostałych, odpowiednio),
  - $GS\big(\mathbf{A}(T_q), \mathbf{A}(\neg T_q)\big) = \frac{\big|\mathbf{A}(T_q)\big|}{|\mathbf{A}|} \cdot Gini\big(\mathbf{A}(T_q)\big) + \frac{\big|\mathbf{A}(\neg T_q)\big|}{|\mathbf{A}|} \cdot Gini\big(\mathbf{A}(\neg T_q)\big)$ jest ważoną sumą współczynników Giniego cięcia p oraz p_i, odpowiednio ($q \in \{p, p_1, \ldots, p_k\}$).

Algorytm 6.2.1: Konstruowanie V-drzewa decyzyjnego

WEJŚCIE: Tablica decyzyjna $\mathbf{A} = (U, A \cup \{d\})$, parametr k należący do liczb naturalnych

WYJŚCIE: *V-drzewo decyzyjne* wyznaczone dla tablicy $\mathbf{A}$

begin

1 Znajdź optymalne cięcie p w tablicy $\mathbf{A}$ i przypisz do wzorca $T_p = TL(p)$ oraz $\neg T_p = TR(p)$

2 Znajdź kolekcję binarnych cięć $p_1, \ldots, p_k$ w tablicy $\mathbf{A}$, weryfikujących cięcie p, najlepszych w sensie wybranej miary jakości (wg procedury Algorytm 6.1.1, str. 92) oraz kolekcję wzorców $VTC(T) = \{T\} \cup \{T_1, ..., T_k\}$ związanych z cięciami weryfikującymi (jeżeli liczba wszystkich wzorców dla danego T jest mniejsza niż k, zbiór może być mniejszy, ale niepusty, ponieważ T zawsze do niego należy)

3 Podziel tablicę $\mathbf{A}$ na dwie tablice $\mathbf{A}(T_p)$ i $\mathbf{A}(\neg T_p)$

4 Przypisz $\mathbf{A}_l = \mathbf{A}(T_p)$ oraz $\mathbf{A}_r = \mathbf{A}(\neg T_p)$

5 Wyznacz wszystkie obiekty z tablicy $\mathbf{A}$, które pasują do wzorca T_p i nie pasują do wzorca T_{p_i} (dla $i \in \{1, ..., k\}$) lub pasują do wzorca $\neg T_p$ i nie pasują do wzorca $\neg T_{p_i}$ (dla $i \in \{1, ..., k\}$) i dołącz te obiekty zarówno do tablicy $\mathbf{A}_l$ jak i $\mathbf{A}_r$ (jeśli jeszcze ich tam nie ma)

6 **jeżeli** *tablice* $\mathbf{A}_l$ *i* $\mathbf{A}_r$ *spełniają warunki stopu* **to**
  zakończ tworzenie drzewa
inaczej
  powtarzaj kroki 1-6 dla wszystkich tablic nie spełniających warunku stopu
end

end

Opis: Algorytm rozpoczyna działanie w węźle zawierającym obiekty wejściowej tablicy decyzyjnej. W kroku pierwszym wyznaczany jest najlepszy podział węzła określony przez cięcie p w sensie przyjętej miary jakości podziału (liczba par obiektów należących do różnych klas decyzyjnych rozróżnianych przez cięcie, zysk informacji lub indeks Giniego, wyliczane według wzorów podanych w Rozdz. 3.1.2). Cięcie główne $p = (b, t)$ definiuje w węźle dwa wzorce T_p i $\neg T_p$. Do T_p przypisywany jest wzorzec lewy ($TL(p) = \{u \in U : b(u) < t\}$ dla atrybutu numerycznego i $TL(p) = \{u \in U : b(u) = t\}$ dla symbolicznego), natomiast do $\neg T_p$ wzorzec prawy ($TR(p) = \{u \in U : b(u) \geq t\}$ i $TR(p) = \{u \in U : b(u) \neq t\}$ odpowiednio). W kroku drugim, przy ustalonym cięciu głównym p wyliczonym w kroku pierwszym, wyznaczane są odpowiednio dobre cięcia weryfikujące to cięcie, według Algorytmu 6.1.1. Kolejny, trzeci krok polega na podzieleniu obiektów węzła na dwie części, przy czym do pierwszej części trafiają obiekty pasujące do wzorca T_p (a więc spełniające warunek $b(u) < t$), do prawej zaś - pasujące

do $\neg T_p$ (spełniające warunek $b(u) \geq t$). Obiekty pierwszej części przypisywane są w kroku czwartym do lewego poddrzewa, natomiast z drugiej części do prawego poddrzewa. W kroku piątym, zarówno do lewego jak i do prawego poddrzewa, dodawane są obiekty, które pasują do części wzorców węzła a nie pasują do reszty wzorców, o ile jeszcze ich tam nie ma. Jeżeli np. do cięcia weryfikującego $p_i = (a, v)$ określonego na atrybucie numerycznym został przypisany w Algorytmie 6.1.1 wzorzec lewy, wówczas dodawanymi obiektami będą te, które spełniają warunek $\{(b(u) < t \wedge a(u) \geq v) \vee (b(u) \geq t \wedge a(u) < v)\}$. Natomiast jeżeli do cięcia p_i przypisany został wzorzec prawy, wówczas dodawane będą obiekty, które spełniają warunek: $\{(b(u) < t \wedge a(u) < v) \vee (b(u) \geq t \wedge a(u) \geq v)\}$. Krok szósty sprawdza, czy lewe jak i prawe poddrzewo spełnia warunek zatrzymania podziału, którym może być obecność w węźle obiektów należących tylko do jednej klasy. Jeżeli tak, budowa drzewa kończy się, w przeciwnym wypadku algorytm rozpoczyna pracę od początku, przy czym dane poddrzewo traktowane jest teraz jako tablica wejściowa.

Zauważmy, że jedynym elementem powyższego algorytmu, który mógłby podwyższyć rząd złożoności czasowej w stosunku do klasycznego algorytmu z Rozdziału 3.1 jest krok 2, w którym wyszukiwana jest kolekcja k binarnych podziałów weryfikujących podział p. Jak zostanie pokazane, krok ten daje się zrealizować w czasie rzędu $O(n \cdot log\ n \cdot m)$, gdzie n jest liczbą obiektów, m liczbą atrybutów warunkowych, a zatem nie powoduje zwiększenia złożoności czasowej algorytmu w stosunku do algorytmu z Rozdziału 3.1.

Łatwo zauważyć, że dla atrybutów symbolicznych, które zwykle mają mało wartości, wyznaczenie najlepszego podziału weryfikującego może być wykonane w czasie $O(n \cdot l)$, gdzie l jest liczbą wartości danego atrybutu symbolicznego.

Klasyfikator skonstruowany za pomocą *V-drzewa decyzyjnego* będzie nazywany *VTree* klasyfikatorem. W zależności od przyjętej miary jakości podziałów (QV_{Disc}, $QV_{Entropy}$, QV_{Gini} wyszczególnia się takie jego rodzaje jak: *VTree-Disc*, *VTree-Entropy*, *VTree-Gini*.

6.3 Klasyfikacja z V-drzewem decyzyjnym

W rozdziale przedstawiono algorytm klasyfikowania obiektu testowego, przy wykorzystaniu drzewa z podziałami weryfikującymi (Algorytm 6.3.1). Załóżmy, że klasyfikujemy obiekt u w węźle, w którym wyszukano optymalne cięcie $c = (a, v)$ oraz rodzinę cięć weryfikujących $c_1, \ldots, c_k$. Niech przez T oznaczony będzie wzorzec generowany przez cięcie c, a przez T_1,...,T_k wzorce odpowiadające cięciom c_1,...,c_k, gdzie dla dowolnego $i \in \{1, \ldots, k\}$ wzorzec $T_i = TL(c_i)$ lub $T_i = TR(c_i)$, zależnie od wzorca wyselekcjonowanego dla danego c_i przez Algorytm 6.2.1. Klasyfikacja odbywa się według Algorytmu 6.3.1.

Algorytm 6.3.1: Klasyfikacja za pomocą V-drzewa

WEJŚCIE: Nowy (testowany) obiekt u, V-drzewo decyzyjne $VT(\mathbf{A})$, wyliczone dla tablicy decyzyjnej $\mathbf{A}$;
$VTC(T) = \{T\} \cup \{T_1, ..., T_k\}$ oznacza kolekcję wzorców dla cięcia optymalnego oraz cięć weryfikujących wyznaczonych przez Algorytm 6.2.1 dla danego węzła w drzewie $VT(\mathbf{A})$

WYJŚCIE: Wartość decyzji dla obiektu u

begin

1 Wstaw obiekt u do korzenia drzewa

2 **jeżeli** *węzeł spełnia warunek stopu* **to**
  zwróć decyzję przypisaną do węzła drzewa i zakończ

3 Przypisz $l_1 :=$ liczba wzorców z $VTC(T)$, do których pasuje u
Przypisz $l_2 := l - l_1$, gdzie $l = card(VTC(T)$

4 **jeżeli** *obiekt u pasuje do wzorca T oraz* $l_1 = l$ **to**
  poślij u do sklasyfikowania przez poddrzewo skonstruowane dla tablicy $\mathbf{A}(T)$, czyli rekurencyjnie wywołaj Algorytm 6.3.1.
  Otrzymaną wartość decyzji oznaczmy przez d_1, **zwróć** d_1 i zakończ.

5 **inaczej jeżeli** *obiekt u nie pasuje do wzorca T oraz* $l_2 = l$ **to**
  poślij u do sklasyfikowania przez poddrzewo skonstruowane dla tablicy $\mathbf{A}(\neg T)$, czyli rekurencyjnie wywołaj Algorytm 6.3.1.
  Otrzymaną wartość decyzji oznaczmy przez d_2, **zwróć** d_2 i zakończ.

6 **inaczej**
  Sklasyfikuj u przez poddrzewo węzła $\mathbf{A}(T)$ otrzymując d_1
  Sklasyfikuj u przez poddrzewo węzła $\mathbf{A}(\neg T)$ otrzymując d_2

7 **jeżeli** $d_1 = d_2$ **to**
  zwróć d_1

8 **inaczej** //Rozstrzyganie konfliktów między d_1 i d_2
  Przypisz p_1:= (rozmiar liścia generującego decyzję d_1)/$|\mathbf{A}|$
  Przypisz p_2:= (rozmiar liścia generującego decyzję d_2)/$|\mathbf{A}|$

9 **jeżeli** $(\frac{l_1}{l} \cdot p_1) > (\frac{l_2}{l} \cdot p_2)$ **to**
  zwróć d_1

10 **inaczej jeżeli** $(\frac{l_1}{l} \cdot p_1) < (\frac{l_2}{l} \cdot p_2)$ **to**
  zwróć d_2
  inaczej //Decyduje cięcie główne

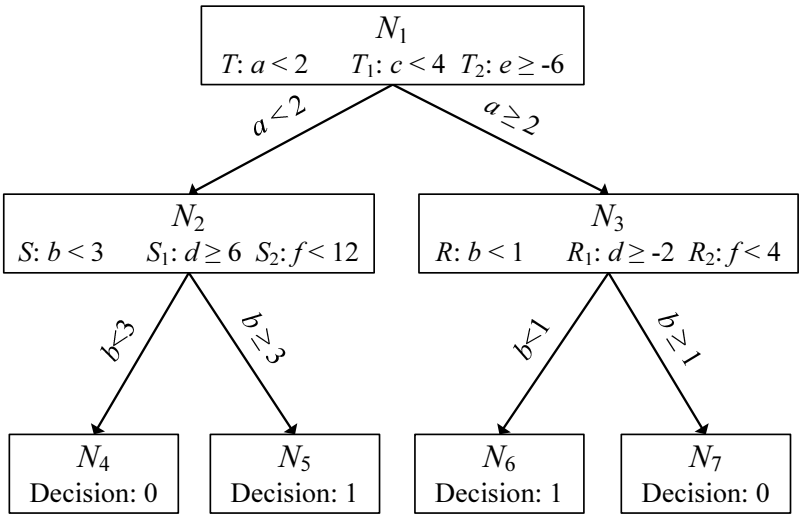
11 **jeżeli** *obiekt u pasuje do wzorca T* **to**
  zwróć d_1

12 **inaczej**
  zwróć d_2

Opis: Algorytm rozpoczyna działanie w węźle, którym na początku jest korzeń drzewa. Sprawdza się tu, czy węzeł jest liściem. Jeżeli tak, to do klasyfikowanego obiektu u przypisuje się etykietę klasy decyzyjnej tego węzła (liścia) i algorytm kończy działanie. W przeciwnym przypadku, w kroku trzecim wyliczana jest liczba wzorców zdefiniowanych dla węzła, do których obiekt u pasuje oraz liczba wzorców, do których obiekt ten nie pasuje. Jeżeli obiekt u pasuje do wszystkich wzorców węzła, w tym do wzorca wyznaczonego przez cięcie główne, wówczas posyłany jest w kroku czwartym do sklasyfikowania przez lewe poddrzewo i przypisywana mu jest decyzja zwracana przez lewe poddrzewo. Jeżeli natomiast obiekt nie pasuje do żadnego wzorca węzła, wówczas jest posyłany do sklasyfikowania przez prawe poddrzewo i przypisywana jest mu decyzja prawego poddrzewa. W przypadku, gdy obiekt u pasuje do części wzorców oraz nie pasuje do pozostałej części wzorców, mamy do czynienia z rozbieżnością wskazań poszczególnych cięć węzła. Aby rozstrzygnąć taki konflikt wyznaczane jest prawdopodobieństwo decyzji z lewego jak i prawego poddrzewa (jako iloraz liczby obiektów w liściach, z których pochodzi decyzja lewego jak i prawego drzewa przez rozmiar całej tablicy treningowej) i określana jest tzw. "siła" decyzji z poszczególnych poddrzew. Siła decyzji wyliczana jest jako iloczyn prawdopodobieństwa decyzji oraz odsetka cięć, które wskazują dane poddrzewo. Jeżeli siła decyzji z lewego poddrzewa jest większa od tej siły prawego poddrzewa, wówczas zwracana jest decyzja lewego poddrzewa. W przeciwnym przypadku zwracana jest decyzja poddrzewa prawego. W przypadku, gdy obie siły są równe, wówczas bierze się pod uwagę wskazanie cięcia głównego.

Celem prezentacji klasyfikacji za pomocą *V-drzewa* (Algorytm 6.3.1), rozważona zostanie tablica decyzyjna $\mathbf{A} = (U, A \cup \{z\})$, taka że $A = \{a, b, c, d, e, f\}$ z atrybutem decyzyjnym z o dwóch wartościach: 0 i 1, co daje dwie klasy decyzyjne: Z_0 i Z_1. Rysunek 6.2 przedstawia *V-drzewo decyzyjne* wyznaczone dla tablicy $\mathbf{A}$ dla $k = 2$ za pomocą Algorytmu 6.2.1. Drzewo to składa się z korzenia N_1, dwóch węzłów wewnętrznych: N_2 i N_3 oraz czterech liści: N_4, N_5, N_6 i N_7. Z węzłem N_1 związany jest wzorzec T dotyczący głównego cięcia w tym węźle oraz dwa wzorce T_1 i T_2 dotyczące cięć weryfikujących. W węźle N_2 zdefiniowany jest wzorzec S związany z cięciem głównym tego węzła oraz dwa wzorce S_1 i S_2 dotyczące cięć weryfikujących. Natomiast do węzła N_3 przypisany jest wzorzec R oparty na cięciu głównym tego węzła oraz dwa wzorce R_1 i R_2 dotyczące cięć weryfikujących go.

Do testowania Algorytmu 6.3.1 wykorzystane zostaną trzy obiekty z Tablicy 6.1. W przykładzie zastosowana będzie prosta metoda rozstrzygania konfliktów między wskazaniami poszczególnych cięć w węźle, oparta na głosowaniu większościowym (gdy obiekt pasuje do części wzorców oraz nie pasuje do pozostałej części wzorców, przypisywana jest mu decyzja pochodząca z poddrzewa, na które wskazuje większa liczba wzorców węzła) [12]. Klasyfikacja wszystkich obiek-



Rysunek 6.2: Przykładowe *V-drzewo decyzyjne.*

tów zaczyna się w węźle N_1.

Obiekt testowy	a	b	c	d	e	f
u_1	1	4	3	-7	-5	14
u_2	3	0	5	-3	-7	5
u_3	1	4	3	-5	-8	10

Tablica 6.1: Tablica obiektów testowych.

Obiekt u_1 pasuje do wzorca T, ponieważ $a(u_1) = 1 < 2$. Jednocześnie obiekt u_1 pasuje do wzorca T_1, ponieważ $c(u_1) = 3 < 4$ oraz pasuje do wzorca T_2, ponieważ $e(u_1) = -5 \geq -6$. Oznacza to, że wzorce T, T_1 i T_2 sugerują, aby obiekt u_1 został sklasyfikowany w węźle N_2. W węźle N_2, obiekt u_1 nie pasuje do wzorca S, ponieważ $b(u_1) = 4 \geq 3$. Ponadto, obiekt u_1 nie pasuje również do wzorca S_1, gdyż $d(u_1) = -7 < 6$ oraz nie pasuje do wzorca S_2, ponieważ $f(u_1) = 14 \geq 12$. Oznacza to, że obiekt u_1 jest kierowany do klasyfikacji przez węzeł N_5 i zostaje sklasyfikowany do klasy decyzyjnej Z_1.

Obiekt u_2 nie pasuje do wzorca T, ponieważ $a(u_2) = 3 \geq 2$. Jednocześnie obiekt u_2 nie pasuje do wzorca T_1, gdyż $c(u_2) = 5 \geq 4$, a także nie pasuje do wzorca T_2, ponieważ $e(u_2) = -7 < -6$. Zatem wzorce T, T_1 i T_2 sugerują, że obiekt u_1 powinien zostać sklasyfikowany w węźle N_3. Zatem obiekt u_2 jest kierowany do klasyfikacji w węźle N_3. W węźle N_3, obiekt u_2 pasuje do wzorca R, ponieważ $b(u_2) = 0 < 1$. Jednocześnie obiekt u_2 nie pasuje do wzorca R_1, gdyż $d(u_2) = -3 < -2$ oraz nie pasuje do wzorca R_2, ponieważ $f(u_2) = 5 \geq 4$. Mamy tutaj do czynienia

z rozstrzyganiem konfliktów, gdyż obiekt u pasuje do jednego z trzech wzorców, a nie pasuje do dwóch pozostałych wzorców węzła. Zatem obiekt u_2 powinien być skierowany do klasyfikacji przez obydwa węzły N_6 i N_7. Niech liczebności węzłów N_1, N_6 i N_7 wynoszą odpowiednio 100, 25 i 40, wówczas siła decyzji pochodzącej z N_6 (Z_1) jest równa $1/3 \cdot 25/100 = 0.08$, natomiast z węzła N_7 (Z_0) jest równa $2/3 \cdot 40/100 = 0.26$. W związku z tym, obiekt u_2 jest zaklasyfikowany do klasy decyzyjnej Z_0.

Obiekt u_3 pasuje do wzorca T, ponieważ $a(u_3) = 1 < 2$. Jednocześnie obiekt u_3 pasuje do wzorca T_1, ponieważ $c(u_3) = 3 < 4$ oraz nie pasuje do wzorca T_2, ponieważ $e(u_3) = -8 < -6$. W związku z pojawiającym się konfliktem, obiekt zostaje posłany do sklasyfikowania przez węzeł N_2 jak i N_3. W węźle N_2, obiekt u_3 nie pasuje do wzorca S, gdyż $b(u_3) = 4 \geq 3$. Jednocześnie obiekt u_3 nie pasuje do wzorca S_1, ponieważ $d(u_3) = -5 < 6$ oraz pasuje do wzorca S_2, ponieważ $f(u_3) = 10 < 12$. Zatem obiekt u_3 zostanie skierowany do klasyfikacji przez dwa węzły N_4 i N_5. Przy liczebnościach węzłów N_4 i N_5 wynoszących odpowiednio 5 i 30, siła decyzji pochodzącej z N_4 (Z_0) wynosi $1/3 \cdot 5/100 = 0.02$, natomiast z węzła N_5 (Z_1) jest równa $2/3 \cdot 30/100 = 0.2$. Do węzła N_2 zwracana jest zatem decyzja Z_1. Natomiast w węźle N_3, obiekt u_3 nie pasuje do wzorca R, gdyż $b(u_3) = 4 \geq 1$. Jednocześnie obiekt u_3 nie pasuje do wzorca R_1, ponieważ $d(u_3) = -5 < (-2)$ oraz nie pasuje do wzorca R_2, ponieważ $f(u_3) = 10 \geq 4$. Zatem obiekt u_3 zostanie skierowany do klasyfikacji przez węzeł N_7, który zwraca do N_3 decyzję Z_0. Siła decyzji otrzymanej dla węzła N_2 jest równa $2/3 \cdot 30/100 = 0.2$, natomiast dla węzła N_3 jest równa $1/3 \cdot 40/100 = 0.13$. W związku z tym, obiekt u_2 jest zaklasyfikowany do klasy decyzyjnej Z_1.

Powyższy algorytm klasyfikowania obiektu w węźle, wykorzystuje jedno poddrzewo tylko w przypadku, gdy wszystkie podziały weryfikujące tak samo klasyfikują obiekt, jak cięcie główne c. W pozostałych przypadkach klasyfikacja jest wykonywana przez obydwa poddrzewa. Następnie rozważane są dwa przypadki. Pierwszy dotyczy sytuacji, gdy obydwa poddrzewa zwróciły tą samą wartość decyzji. Wtedy ta wartość jest zwracana jako decyzja danego węzła. Natomiast drugi przypadek dotyczy sytuacji, gdy jedno z poddrzew zwróciło jedną wartość decyzji, a drugie poddrzewo inną. Wtedy z danego węzła jest zwracana decyzja pochodząca z tego poddrzewa, które wiąże się z większą siłą decyzji. Taka metoda stanowi prosty sposób na rozstrzyganie konfliktów między decyzjami wygenerowanymi przez dwa drzewa.

Rozdział 7

Metoda IV: Definiowanie odległości ontologicznej i jej zastosowanie do konstrukcji klasyfikatorów metodą k-NN

Zawartość

Jednym z problemów występujących w procesie odkrywania wiedzy ze zbiorów danych jest złożoność procesów zachodzących w rzeczywistym świecie, obecność bezpośrednich i pośrednich powiązań oraz interakcji pomiędzy obiektami biorącymi w nich udział. Klasyczne metody modelowania bazujące na danych pochodzących z czujników nie umożliwiają badania danych na wielu poziomach abstrakcji. Jest to wynikiem oddalenia semantycznego złożonych pojęć od danych sensorowych (patrz Rozdz. 2.2).

Przy budowie klasyfikatorów aproksymujących złożone pojęcia może wystąpić potrzeba oceny odległości lub podobieństwa dwóch obiektów podobnego typu, takich jak np. pacjenci. Ogólnie, problem definiowania odległości lub podobieństwa jest ciągle jednym z największych wyzwań eksploracji danych. Istniejące metody definiowania relacji podobieństwa oparte są zwykle na budowaniu funkcji odległości w oparciu o proste strategie łączenia lokalnych podobieństw porównywanych elementów (patrz literaturę w [20]). Optymalizacja ustalonej formuły odległości jest wykonywana poprzez strojenie parametrów lokalnych podobieństw oraz parametrów ich łączenia. Główną trudnością aproksymacji funkcji odległości jest zatem

dobór tych lokalnych podobieństw oraz sposobu ich agregacji. Tymczasem, zgodnie z wiedzą dziedzinową zwykle jest wiele różnych aspektów podobieństwa pomiędzy porównywanymi elementami. Każdy z tych aspektów powinien być rozpatrywany w specyficzny sposób, zgodny z wiedzą dziedzinową. Ponadto agregacja różnych aspektów podobieństwa w globalne podobieństwo czy odległość, także powinna być wykonana w sposób wynikający z wiedzy dziedzinowej. Dlatego w rozprawie zaproponowano metodę definiowania odległości opartą na funkcji podobieństwa wykorzystującą wiedzę dziedzinową wyrażoną w postaci ontologii pojęć.

Zaproponowana, na potrzeby eksploracji rzeczywistego zbioru danych medycznych, funkcja podobieństwa ma umożliwiać porównywanie pacjentów pod kątem stopnia nasilenia choroby wieńcowej, a co za tym idzie ryzyka występowania niebezpiecznych dla zdrowia i życia następstw. Im bardziej zaawansowana choroba, tym większe ryzyko tzw. incydentów sercowych (groźne zaburzenia rytmu, ostre stany niedokrwienia mięśnia sercowego czy nagły zgon sercowy NZS). Celem oceny efektywności takiego podejścia wyznaczoną odległość ontologiczną zastosowano w zadaniu klasyfikacji. W tym podejściu klasy decyzyjne oznaczają stopień nasilenia stabilnej choroby wieńcowej, gdzie 0 oznacza najmniej nasiloną chorobę, 1 - chorobę jednonaczyniową, 2 - dwunaczyniową i 3 - chorobę trójnaczyniową, czyli o największym stopniu nasilenia.

W pierwszym etapie konstrukcji funkcji podobieństwa definiowana jest hierarchiczna ontologia zawierająca pojęcia dotyczące choroby niedokrwiennej serca. Na najniższym poziomie znajdują się atrybuty sensorowe (pochodzące bezpośrednio ze zbiorów danych), dobrane z całego zbioru danych w taki sposób, aby odpowiadały uznanym czynnikom prognostycznym NZS. Następnie na każdym poziomie ontologii określana jest ważność (znaczenie) pojęcia w odniesieniu do pojęcia nadrzędnego poprzez wskazanie wagi. Wagi dobierane są arbitralnie przez eksperta dziedzinowego, jako wartości liczbowe z przedziału $(0, 1)$. Na potrzeby eksperymentów utworzono ontologię jak na Rys. 7.1, w której wagi pojęć zostały wskazane przez eksperta. Dodatkowo w doświadczeniach wykorzystano tę samą ontologię, jednak z wagami pojęć dobranymi metodą Monte Carlo. Kolejny krok polega na wyspecyfikowaniu algorytmu obliczania funkcji podobieństwa pomiędzy obiektami z wykorzystaniem ontologii i wag jej pojęć. W następnym etapie na podstawie podobieństwa semantycznego pomiędzy pacjentami budowany jest klasyfikator.

7.1 Budowa ontologii

Do wyznaczenia odległości ontologicznej wymagane jest zdefiniowanie ontologii pojęć należących do pojęcia określającego problem decyzyjny. Ontologie stanowią opis fragmentu świata służący do reprezentowania i przetwarzania wiedzy.

Według podejścia zaproponowanego w [102] proces budowania ontologii składa

się z kilku czynności wymienionych poniżej.

1. Ustalenie domeny oraz zasięgu ontologii - określenie jakiego wycinka modelowanego świata będzie dotyczyć.
2. Wykorzystanie istniejących ontologii.
3. Wyszczególnienie najważniejszych terminów w projektowanej ontologii.
4. Definiowanie klas i ich uporządkowanie w hierarchiczne struktury na kształt drzewa (nadklasy i podklasy).
5. Zdefiniowanie własności klas.
6. Definiowanie cech własności klas, takich jak np. ich dziedziny, czyli dopuszczalne zbiory wartości.
7. Tworzenie wystąpień klas (instancji klas).

Opierając się częściowo na powyższej metodologii utworzono ontologię O_{CNS} pojęcia *stabilna choroba wieńcowa*. Kolejne etapy konstrukcji ontologii obejmowały:

- Ustalenie dziedziny nauk medycznych jako domeny ontologii oraz kardiologii jako jej zasięgu;
- Wyszczególnienie terminów, które wskazują na stopień zaawansowania choroby niedokrwiennej serca, takich jak: zmiany w badaniu podmiotowym pacjenta, zmiany w badaniach dodatkowych, zagrożenia epidemiologiczne, obecność chorób współistniejących, zmiany w badaniach elektrofizjologicznych, odchylenia w badaniach laboratoryjnych;
- Zdefiniowanie następujących klas ontologii: CNS, Badanie podmiotowe, Badania dodatkowe, Epidemiologia, Choroby współistniejące, Badania elektrofizjologiczne, Badania laboratoryjne, EKG, HRV, QT, Tachykardia, ST. Przy porządkowaniu klas w hierarchiczne struktury zastosowano podejście "góra-dół" (ang. *top-down*), w którym zaczyna się od najbardziej ogólnych pojęć i przechodzi kolejno do ich uszczegóławiania;
- Określenie znaczenia każdego pojęcia w odniesieniu do pojęcia nadrzędnego poprzez wprowadzenie wagi jako własności klasy;
- Wskazanie dziedziny wag jako liczb rzeczywistych z przedziału $(0, 1)$. Wagi zostały dobrane arbitralnie przez eksperta dziedzinowego, w taki sposób, że suma wag pojęć podrzędnych (podklas) odpowiada wadze pojęcia nadrzędnego (nadklasy). Można zatem powiedzieć o zjawisku rozpływania się wag, od pojęcia najbardziej ogólnego do najbardziej szczegółowych;
- Zdefiniowanie instancji poszczególnych klas w postaci uznanych czynników prognostycznych nagłego zgonu sercowego NZS [117, 55, 57, 3]. Wybrane

czynniki ryzyka przedstawiają tabele: 7.1 oraz 7.2. W ontologii CNS wykorzystano 19 czynników ryzyka, którym przypisano wagi zależnie od znaczenia czynnika w obrębie pojęcia, do którego należą. Instancje ontologii odpowiadają atrybutom zbioru danych.

Badanie	**Czynnik ryzyka**	**Opis**
Epidemiologia	Wiek	>65 rż.
	Płeć	Mężczyźni:Kobiety = 4:1
	Używki	tytoń (czynnik ryzyka)
Choroby współistniejące	Cukrzyca	niekorzystne prognostycznie
	Nadciśnienie	niekorzystne prognostycznie
	Zawał serca	niekorzystne prognostycznie (ew. liczba przebytych)
	Miażdżyca tętnic kkd	niekorzystne prognostycznie
	Udar mózgu	niekorzystne prognostycznie

Tablica 7.1: Czynniki prognostyczne NZS w badaniu podmiotowym.

W literaturze brak określenia wymaganej liczby czynników ryzyka – im więcej czynników, tym większe ryzyko tzw. incydentów sercowych (groźne zaburzenia rytmu, ostre stany niedokrwienia mięśnia sercowego, nagły zgon sercowy NZS).

Utworzoną ontologię pojęć choroby niedokrwiennej serca (CNS) przedstawia Rys. 7.1. Najniższy poziom stanowią instancje klas. Zaproponowana ontologia zawiera tylko wybrane pojęcia, których instancje były dostępne w zbiorach danych, jednak może być z łatwością rozszerzona poprzez dodawanie pojęć oraz instancji.

7.2 Wyznaczanie odległości ontologicznej

Na podstawie ontologii wyznaczana jest odległość między dwoma obiektami przynależącymi do danego pojęcia, dla którego zbudowana jest ontologia (każde pojęcie w tej ontologii opisuje zróżnicowanie pacjentów). Odległość ontologiczna utworzona na potrzeby rozprawy ma odpowiadać na pytanie: "jak podobni (niepodobni) są do siebie pacjenci z chorobą niedokrwienną serca?".

W klasycznych technikach opartych na odległości, stosuje się takie miary jak odległość Euklidesa, czy ogólnie odległość Minkowskiego (*p-norma*) wyrażone wzorami 3.7 i 3.8 (Rozdział 3.2). Odległości te jednak uwzględniają tylko dane z poziomu sensorów, nie biorąc pod uwagę zależności i powiązań między pojęciami na

Badanie	Czynnik ryzyka	Opis
Badania elektrofizjologiczne (EKG)	Zmiany ST	obniżenia>0,5 mm, uniesienie>1 mm
	Zmiany QTc	norma <440 ms
	Tachykardia	spoczynkowy HR>70/min, liczba wystąpień
	Arytmia	0-5 (skala Lown), gdzie 0 - brak przedwczesnych pobudzeń komorowych
Testy laboratoryjne	Troponina I	norma<0,01 ug/l
	LDL	norma<3,5 mmol/l
	CRP	norma<10 mg/l

Tablica 7.2: Czynniki prognostyczne NZS w badaniach dodatkowych.

wyższym poziomie abstrakcji. Dodatkowo wymagają, aby atrybuty były numeryczne.

Odległość ontologiczna natomiast uwzględnia hierarchię i znaczenie pojęć należących do ontologii. Zaproponowana w rozprawie metodologia wyznaczania odległości ontologicznej obejmuje dwa etapy. W pierwszym wyznaczane są odległości między wartościami parametrów wskazywanych przez czujniki, a więc na najniższym poziomie ontologii. W kolejnym kroku określana jest odległość na poziomie każdego z pojęć wyższego poziomu z uwzględnieniem ich znaczenia wyrażonego poprzez wagi.

Funkcja podobieństwa pomiędzy dwoma obiektami u_i i u_j względem numerycznego atrybutu sensorowego a znajdującego się na najniższym poziomie ontologii jest wyznaczana zgodnie ze wzorem (7.1) [166]:

$$d_{num}(u_i, u_j, a) = \frac{|a(u_i) - a(u_j)|}{R_a} \qquad \text{dla } i, j \in \{1, ..., n\} \tag{7.1}$$

gdzie n to liczba obiektów, natomiast R_a to rozstęp wartości atrybutu. Rozstęp może być wyznaczany jako odległość między maksymalną i minimalną wartością atrybutu w badanym zbiorze lub zakres wartości znany z wiedzy dziedzinowej. Ze względu na brak ściśle określonych wartości granicznych niektórych czynników ryzyka NZS w rozprawie zastosowano podejście pierwsze.

Funkcja podobieństwa natomiast względem symbolicznego (nienumerycznego) atrybutu sensorowego a wyznaczana jest z wykorzystaniem metody VDM (ang. *value difference metric*) [145], w której odległość wyznaczana jest według wzoru

(7.2):

$$d_{symb}(u_i, u_j, a) = \sum_{d_c \in D} |P(dec = d_c | a(u_i) = v) - P(dec = d_c | a(u_j) = w)| \quad (7.2)$$

gdzie D to zbiór klas decyzyjnych, P to rozkład prawdopodobieństwa wartości decyzji (wzór 7.3), $v, w \in V_a$.

$$P(dec = d_c | a(u) = v) = \frac{|\{u \in U : dec(o) = d_c \wedge a(u) = v\}|}{|\{u \in U : a(u) = v\}|} \quad (7.3)$$

Natomiast funkcja podobieństwa pomiędzy dwoma obiektami u_i i u_j względem pojęcia C znajdującego się na wyższym poziomie ontologii definiowana jest zgodnie ze wzorem (7.4):

$$d_{onto}(u_i, u_j, C) = \begin{cases} \sum_{s \in S} w_s \cdot d_{Onto}(u_i, u_j, C_s) & \\ \sum_{a \in A} w_s \cdot d_{num}(u_i, u_j, a) & \textit{dla atr. numerycznych} \\ \sum_{a \in A} w_s \cdot d_{symb}(u_i, u_j, a) & \textit{dla atr. symbolicznych} \end{cases} \quad (7.4)$$

gdzie S to zbiór pojęć podrzędnych przynależących do pojęcia C, w_s - waga pojęcia podrzędnego, natomiast C_s to pojęcie podrzędne dla C (na najniższym poziomie ontologii odpowiada atrybutowi).

7.3 Odległość ontologiczna jako metryka

Odległość ontologiczna zaproponowana w rozprawie spełnia warunki miary odległości wymienione w Rozdz. 3.2. Własności pierwsza i druga (aksjomat tożsamości i aksjomat symetrii) wynikają bezpośrednio z własności wartości bezwzględnej. Udowodniona zostanie własność trzecia, tj. spełnianie nierówności trójkąta.

Twierdzenie. *Niech d_1, d_2 będą metrykami w przestrzeniach 1-wymiarowych. Wtedy $D = \sum_{i=1}^{k} w_i \cdot d_1 + \sum_{i=k+1}^{n} w_i \cdot d_2$ jest metryką w przestrzeni n-wymiarowej dla dowolnych wartości wag $w_i \geq 0$, dla i=1,2,...,n, takich że $\sum_{i=1}^{n} w_i > 0$.*

Dowód. *Dowód warunku trójkąta indukcyjnie ze względu na wymiar przestrzeni, gdzie wymiar przestrzeni to liczba atrybutów warunkowych w systemie decyzyjnym, czyli liczba współrzędnych wektora opisującego poszczególne obiekty. Wśród wszystkich atrybutów k jest numerycznych, a n-k+1 – symbolicznych (lub odwrotnie).*

1^o W przestrzeni 1-wymiarowej $D = w \cdot d_1$ albo $D = w \cdot d_2$, w>0.
Ogólnie: $D = w \cdot d_p$, gdzie $p \in \{1, 2\}$. Z założenia o d_1 i d_2 mamy:

$\forall x, y, z\ d_p(x, y) + d_p(y, z) \geq d_p(x, z)$, a więc także
$\forall x, y, z\ w \cdot d_p(x, y) + w \cdot d_p(y, z) \geq w \cdot d_p(x, z)$ dla w>0.
Stąd $\forall x, y, z\ D(x, y) + D(y, z) \geq D(x, z)$

2^o Załóżmy, że D spełnia warunek trójkąta w przestrzeni m-wymiarowej, czyli:

$$\forall x, y, z \quad \forall_{\substack{w_1,w_2,\ldots,w_m \geq 0 \\ w_1+w_2+\ldots+w_m>0}} \quad \sum_{i=1}^{k} w_i \cdot d_1(x, y) \quad + \quad \sum_{i=k+1}^{m} w_i \cdot d_2(x, y) \quad +$$
$$\sum_{i=1}^{k} w_i \cdot d_1(y, z) + \sum_{i=k+1}^{m} w_i \cdot d_2(y, z) \ \geq \ \sum_{i=1}^{k} w_i \cdot d_1(x, z) + \sum_{i=k+1}^{m} w_i \cdot d_2(x, z)$$

W przestrzeni m+1 wymiarowej będzie:

$$(*)\ D(x, y) + D(y, z) = \sum_{i=1}^{k} v_i \cdot d_1(x, y) + \sum_{i=k+1}^{m} v_i \cdot d_2(x, y) + v_{m+1} \cdot d_p(x, y) +$$
$$\sum_{i=1}^{k} v_i \cdot d_1(y, z) + \sum_{i=k+1}^{m} v_i \cdot d_2(y, z) + v_{m+1} \cdot d_p(y, z),$$ gdzie p=1 lub p=2.

Możemy napisać

$$(*) \geq \sum_{i=1}^{k} v_i \cdot d_1(x, z) + \sum_{i=k+1}^{m} v_i \cdot d_2(x, z) + v_{m+1} \cdot d_p(x, y) + v_{m+1} \cdot d_p(y, z)\ (**)$$

dzięki 2^o, gdy nie wszystkie $v_i = 0$, i=1,2,...,m. A jeśli wszystkie $v_i = 0$, i=1,2,...,m to nierówność jest oczywista, bo zachodzi równość. Dalej

$$(**) \geq \sum_{i=1}^{k} v_i \cdot d_1(x, z) + \sum_{i=k+1}^{m} v_i \cdot d_2(x, z) + v_{m+1} \cdot d_p(x, z)$$

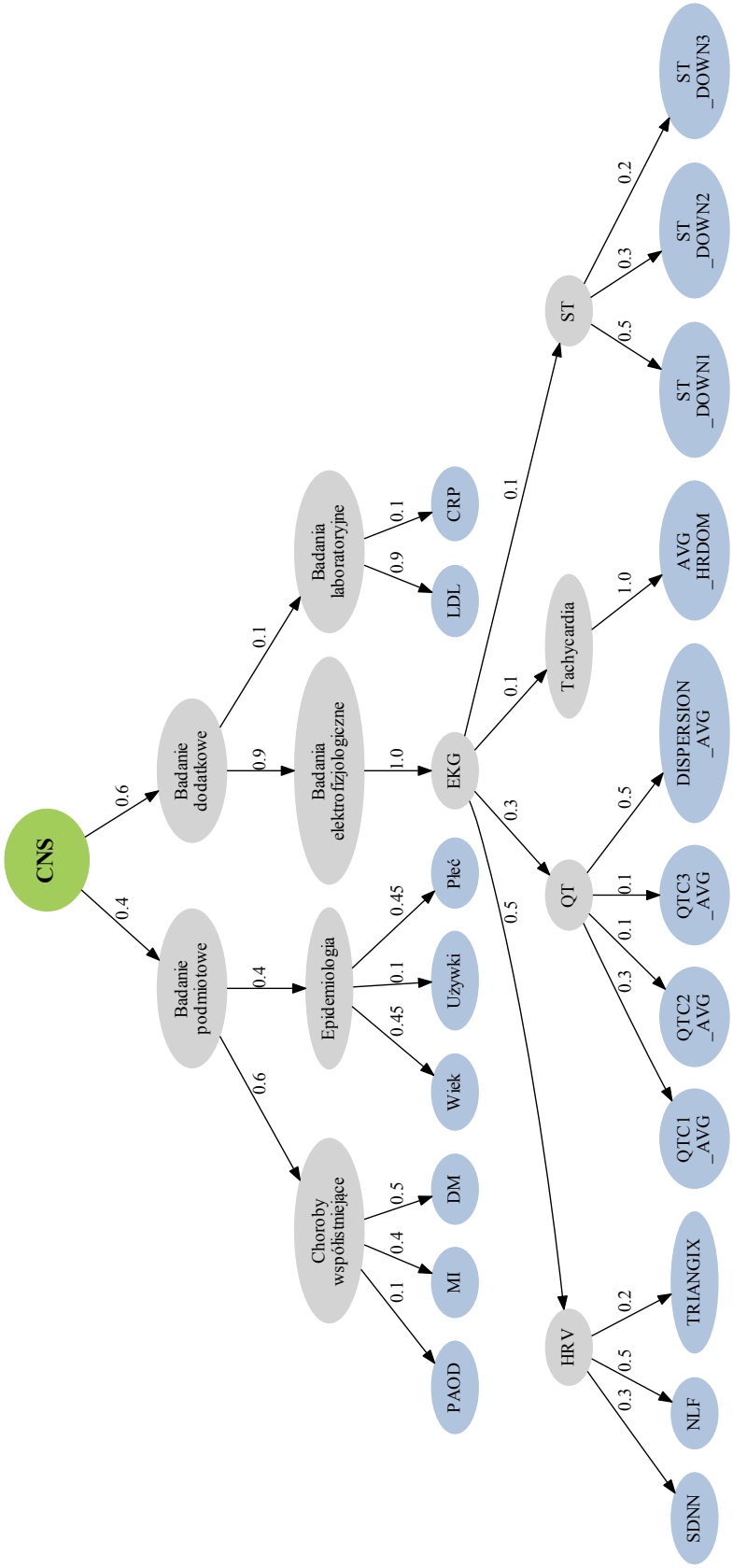
dzięki 1^o, gdy $v_{n+1} \neq 0$, i=1,2,...,m. A jeśli $v_{m+1} = 0$ to nierówność jest oczywista, bo zachodzi równość, c.k.d.

Odległość ontologiczna zdefiniowana rekurencyjnie wzorem (7.4) może być przedstawiona w następujący sposób:

$$d_{onto}(u_i, u_j, C) = \sum_{a \in A_{num}} w_a \cdot d_{num}(u_1, u_j, a_{num}) + \sum_{a \in A_{symb}} w_a \cdot d_{symb}(u_i, u_j, a_{symb}) \tag{7.5}$$

gdzie w_a są wagami wyznaczonymi zgodnie z opisem przedstawionym w tym podrozdziale. Stąd na podstawie przedstawionego twierdzenia i jego dowodu wnioskujemy o spełnialności warunku trójkąta przez odległość ontologiczną.

Przy wykorzystaniu odległości ontologicznej, przeprowadzono eksperymenty z klasyfikatorami metodą k-NN, których wyniki opisano w Rozdziale 9.5.



Rysunek 7.1: Ontologia CNS z wagami wskazanymi przez eksperta.

Rozdział 8

Metoda V: Opis wpływu czynnika modyfikującego percepcję w oparciu o modele klasyfikacji

Zawartość

Jednym z aspektów eksploracji danych jest zdobywanie wiedzy o otaczającym świecie poprzez nadawanie znaczeń otrzymanym wrażeniom, czyli informacjom dostarczonym przez sensory. Sposób, w jaki postrzegamy otoczenie i interpretujemy rzeczywistość istotnie warunkuje identyfikację obiektów i ich klasyfikację, przez co wpływa na podejmowanie decyzji.

Klasyfikatory w oparciu o dostępne cechy (atrybuty warunkowe) charakteryzujące badane obiekty, opisują wartość atrybutu decyzyjnego i mogą być traktowane jako narzędzie do aproksymacji pojęć (klas decyzyjnych) (patrz np. [107, 20, 14, 16]). Dzięki klasyfikatorom możliwa jest zatem percepcja obiektu testowego w kontekście informacji o nim związanych z wartością atrybutu decyzyjnego.

Percepcja definiowana jest zwykle jako proces rozpoznawania, organizowania i interpretacji informacji [162, 127]. Percepcja rozumiana jest tutaj jako sposób postrzegania świata, który prowadzi do ukształtowania się subiektywnego obrazu rze-

czywistości, na podstawie którego podejmowane są decyzje. Na proces postrzegania wpływa wiele czynników, takich jak rodzaj bodźca, sensora, sposób przedstawienia informacji, wcześniejsze doświadczenia obserwatora czy kontekst. W związku z tym pozyskane informacje wpływają na formowanie obrazu rzeczywistości.

Dla potrzeb tej rozprawy pojęcie percepcji ograniczone zostanie do interpretowania informacji pochodzących z systemów diagnostycznych. Rozpatrywana jest percepcja związana z tematyką medyczną, a dokładnie problemem zwężeń tętnic wieńcowych u chorych ze stabilną chorobą niedokrwienną serca CNS (patrz Dodatek A). Zbadany zostanie wpływ czynników "zakłócających" na postrzeganie zwężeń w naczyniach krwionośnych. Pokazane zostanie, że niektóre czynniki "zakłócające" mogą być zarządzane za pomocą algorytmów eksploracji danych opartych na znanych statystykach w połączeniu z krzyżowymi regułami decyzyjnymi.

8.1 Percepcja a klasyfikacja

Istnieje szereg czynników, które mogą modyfikować zachowanie badanego obiektu i zmieniać sposób jego percepcji (postrzegania), pomimo że generalnie stan obiektu się nie zmienia w sensie przynależności do pewnego pojęcia. Interesująca jest w takiej sytuacji zmiana, która prowadzi do zachowania mającego korzystny wpływ na obserwowany obiekt. Modyfikacja może prowadzić również do negatywnych dla obiektu skutków. Bez względu na efekt percepcja informacji na temat badanego obiektu jest zakłócona, natomiast informacja na temat sposobu modyfikacji zachowania może zostać wykorzystana do uzyskania pożądanego zachowania obiektu.

Czasami na czynnik zakłócający nie mamy wpływu (środowisko, pogoda), ale czasem można celowo go wprowadzać lub nawet nim zarządzać, jeśli wiemy jak to robić (leki, rodzaj terapii, dieta). Przykładem czynnika modyfikującego percepcję może być farmakoterapia stosowana u pacjentów w danym schorzeniu. W wyniku zastosowanego leczenia, pacjenci przynależący do pewnego pojęcia mogą zachowywać się - w sensie procesu postrzegania - tak jak pacjenci nie otrzymujący leczenia, a należący do innego (przeciwstawnego) pojęcia.

Metody opisu i rozróżnienia takich grup pacjentów umożliwiłyby wyselekcjonowanie obiektów, u których czynnik zakłócający daje korzystne efekty. Na polu medycyny taka wiedza wskazałaby pacjentów, u których warto stosować daną farmakoterapię, a u których nie.

Jeżeli założymy, że percepcja realizowana jest za pomocą klasyfikatora, to oznacza, że w przypadku pojawienia się czynnika zakłócającego percepcję, klasyfikator musi być przekonstruowany. W takiej sytuacji można budować klasyfikator wykorzystujący dodatkowy atrybut warunkowy reprezentujący informację o występowaniu czynnika zakłócającego. Dzięki temu klasyfikator może działać skutecznie zarówno w sytuacji, gdy czynnik występuje lub w sytuacji, gdy nie występuje.

Można także konstruować dwa klasyfikatory:

1. Pierwszy dla sytuacji, gdy nie ma czynnika zakłócającego.
2. Drugi dla sytuacji, gdy czynnik zakłócający jest obecny.

Osobnym zagadnieniem jest jednak pytanie, w jaki sposób czynnik zakłócający zmienia percepcję obiektów testowych za pomocą klasyfikatora i czy zmiana ta dotyczy w taki sam sposób wszystkich obiektów testowych. Jeśli udałoby się poznać reguły tych zmian, to można by było wykorzystać czynnik zakłócający do sterowania percepcją obiektu i pośrednio jego zachowaniem.

Z obserwacji praktycznych wiadomo bowiem, że często czynnik zakłócający może powodować różne zmiany percepcji obserwowanego obiektu. Dla przykładu rozpatrzmy sytuację, gdy percepcja dotyczy stanu pacjenta opisanego przez cztery stany A, B, C, D związane ze stopniem zaawansowania choroby, która może powodować zagrożenie życia. Przy czym stan A oznacza najmniejsze zaawansowanie choroby, a D największe. Załóżmy także, że percepcja oparta jest na atrybutach opisujących wyniki badań objawowych, które opisują aktualny poziom stanu zagrożenia życia, takich jak np. sygnał EKG. Niech czynnikiem zakłócającym percepcję będzie podanie leku Z (w głównym problemie rozprawy jest to lek zileuton o działaniu przeciwzapalnym). Niech K_1 i K_2 są klasyfikatorami, które zostały skonstruowane w celu predykcji stanu A, B, C lub D na danych treningowych odpowiednio bez i z czynnikiem zakłócającym. Jeśli klasyfikator K_1 danego pacjenta P_1 sklasyfikuje do stanu B, to okazuje się, że klasyfikator K_2 może sklasyfikować P_1 zarówno do stanu A, B, C jak i D. Jest tak dlatego, że czynnik zakłócający percepcję powoduje przekłamanie obrazu pacjenta, który czasami postrzegany jest jakby miał bardziej zaawansowaną chorobę niż ma naprawdę, a czasem jakby miał mniej zaawansowaną chorobę. Przy czym to zafałszowanie obrazu choroby dotyczy nie tyle rzeczywistego zaawansowania choroby, ale aktualnego stopnia zagrożenia życia, które ta choroba powoduje. Gdyby zatem można było przewidywać jaką zmianę percepcji choroby spowoduje lek Z u danego pacjenta, można by było wykorzystać to do doraźnego wspomagania leczenia (np. w drodze do szpitala, w oczekiwaniu na zabieg itd.).

Standardowa metoda tworzenia klasyfikatora jednak nie jest przydatna do przewidywania zmian percepcji spowodowanych czynnikiem zakłócającym, gdyż nie można skonstruować atrybutu decyzyjnego reprezentującego zmianę percepcji choroby dla danego pacjenta. Wymagałoby to bowiem eksperymentów z udziałem pacjentów, którzy musieliby być leczeni zarówno bez jak i z użyciem leku Z, co wydaje się być trudne technicznie i nieetyczne, bo może generować ryzyko mniej efektywnego leczenia (leczenie rozciągnięte w czasie).

Rozpatrywany problem był badany w pracy [24] dotyczącej wpływu zastosowanej metody leczenia chirurgicznego nowotworów gardła i krtani na przeżywalność

pacjentów. Prace te doprowadziły do odkrycia metody eksploracji danych medycznych, która wylicza tzw. *reguły krzyżowe*. Każda reguła krzyżowa, dla danej grupy pacjentów opisanej przez poprzednik reguły, podaje w następniku prawdopodobieństwo sukcesu leczenia (przeżycia pacjenta) w przypadku obydwu badanych metod leczenia.

Tego rodzaju reguły można by było zaproponować jako rozwiązanie problemu przewidywania zmian percepcji. W tym celu w następniku reguł pojawiłyby się informacje o percepcji, zarówno bez, jak i z czynnikiem zakłócającym. Jednak metoda z [24] działa w oparciu o dane z atrybutami symbolicznymi. Ograniczenie to wynika z potrzeby liczenia specyficznych reduktów względem decyzji złożonej, które w publikacji [24] są liczone dla danych symbolicznych. Tymczasem w wielu zastosowaniach trzeba analizować dane z atrybutami numerycznymi lub mieszanymi (symboliczne i numeryczne). Potrzebne są zatem reguły krzyżowe, które w poprzedniku zamiast konkretnych wartości atrybutów, będą mogły mieć przedziały wartości.

8.2 Metoda mierzenia stopnia wpływu czynnika zakłócenia procesu

W rozprawie zaproponowano metodę wyznaczania wpływu czynnika modyfikującego percepcję opartą na tzw. drzewie wpływu. *Drzewo wpływu*, nazywane także *I-drzewem* lub *ITree* (ang. *impact tree*) jest drzewem binarnym, tak skonstruowanym, że w jego liściach znajdują się opisy grup pacjentów (wzorce), którym w podobny sposób zmienia się percepcja choroby po zastosowaniu czynnika zakłócającego. W niektórych liściach obserwuje się dużą zmianę percepcji, a w niektórych małą. W niektórych liściach zmiana jest pozytywna, a w innych negatywna. Dzięki wzorcom przypisanym do liści można określić zmianę percepcji obiektów poddanych działaniu czynnika zakłócającego, nazywanego inaczej modyfikatorem (w języku angielskim dobrym odpowiednikiem jest słowo *confounder*). Każdy zatem liść tego drzewa dostarcza jednej reguły krzyżowej, które w rozprawie będą nazwane *regułami krzyżowymi zmian percepcji*. Reguły te mogą być zaproponowane jako rozwiązanie problemu przewidywania zmiany percepcji. W tym celu, w następniku reguły powinny znaleźć się informacje o postrzeganiu obiektów, zarówno bez i z czynnikiem zakłócającym.

8.2.1 Reguły krzyżowe zmian percepcji

Zaproponowane *reguły krzyżowe zmian percepcji* stanowią rodzaj reguł decyzyjnych. Reguły decyzyjne mają postać implikacji (wzór 8.1), w której po lewej stronie znajdują się warunki (przesłanki) wyrażone pewną formułą logiczną, po prawej

- wniosek (teza):

$$\text{JEŻELI warunki TO wniosek} \tag{8.1}$$

Część warunkowa może przyjmować postać koniunkcji warunków elementarnych i jest reprezentowana w postaci przedstawionej wzorem (8.2):

$$warunki = w_1 \wedge w_2 \wedge ... \wedge w_k \tag{8.2}$$

gdzie k jest liczbą wykorzystanych warunków i określa długość reguły.

W klasycznych regułach decyzyjnych tezą jest przynależność do ustalonej klasy decyzyjnej, w regułach krzyżowych natomiast opis grup obiektów znajdujących się w różnych warunkach. W [24] deskryptory opisują prawdopodobieństwo zajścia pewnego zdarzenia w różnych warunkach, jak we wzorze (8.3):

$$T = wr \wedge E = 0 \wedge S = 1 \Rightarrow \left| \begin{array}{l} P(\text{success after radical}) \quad = p_1 \\ P(\text{success after modified}) = p_2 \end{array} \right. \tag{8.3}$$

gdzie T – treatment, wr – with radiotherapy, E – Extracapsular spread, S – Stage, $(T, E, S) \in A$ (A to zbiór atrybutów), p_1, p_2 - prawdopodobieństwo sukcesu terapeutycznego u pacjentów poddanych odpowiednio terapii radykalnej oraz terapii zmodyfikowanej.

Proponowane natomiast w rozprawie *reguły krzyżowe zmian percepcji* mają postać przestawioną wzorem (8.4):

$$warunki \Rightarrow \left| \begin{array}{l} E(dec \mid Modyfikator = y_1) = x_1 \\ E(dec \mid Modyfikator = y_2) = x_2 \end{array} \right. \tag{8.4}$$

gdzie dec to decyzja, y_1, y_2 to rodzaje czynnika zakłócającego (modyfikatora), x_1, x_2 - warunkowa wartość oczekiwana decyzji w obecności poszczególnych modyfikatorów w badanej grupie obiektów wyznaczona według standardowego wzoru (8.5):

$$\begin{aligned} E(dec \mid Modyfikator = y) = \sum_{d \in dec} d \cdot P(dec = d \mid Modyfikator = y) = \\ \sum_{d \in dec} d \cdot \frac{P(dec = d, Modyfikator = y)}{P(Modyfikator = y)} \end{aligned} \tag{8.5}$$

Każda reguła krzyżowa opisuje więc wartość oczekiwaną decyzji w grupie poddanej działaniu modyfikatora y_1 (równą x_1) oraz w grupie poddanej działaniu modyfikatora y_2 (wynoszącą x_2). W szczególnym przypadku jeden z modyfikatorów może nie mieć wpływu na rozpatrywany problem decyzyjny. Przykładowo dla modyfikatora, którym jest farmakoterapia będzie to podanie placebo. Taki dobór wartości modyfikatora umożliwia porównanie grupy poddanej działaniu tego czynnika z grupą niepoddaną jego wpływowi [34].

Dla rozpatrywanego problemu choroby niedokrwiennej serca niech atrybutem decyzyjnym będzie liczba stenoz naczyń wieńcowych (S) określona na podstawie koronarografii, czynnikiem modyfikującym - terapia dwoma specyfikami: lekiem Z (zileuton) oraz placebo P (w praktyce oznaczająca brak dodatkowego leczenia). Wtedy reguła może przyjmować następującą postać:

$$a = 1 \wedge b > 2 \Rightarrow \left| \begin{array}{l} E(S \mid terapia = P) = 0 \\ E(S \mid terapia = Z) = 2 \end{array} \right. \tag{8.6}$$

gdzie a, b to atrybuty warunkowe. Zatem przeciętna liczba stenoz w grupie leczonej lekiem Z wynosi 2, w grupie nieleczonej (otrzymującej placebo) wynosi 0. W praktyce atrybuty a i b mogą być parametrami zapisu EKG, takimi jak liczba przedwczesnych pobudzeń komorowych serca czy czas trwania odstępu QT. Niech wzorzec ($a = 1 \wedge b > 2$) oznacza obraz stabilny w zapisie EKG (bez niebezpieczeństwa). Wówczas reguła przedstawiona wzorem (8.6) oznacza, że dodatkowa terapia powoduje znaczne przekłamanie rzeczywistego obrazu określonego na podstawie koronarografii w przypadku powyższego wzorca. Pomimo przeciętnie dwóch zwężonych naczyń ($x_2 = 2$), pacjenci poddani leczeniu lekiem Z mają zapis EKG nieodróżnialny od krzywej EKG pacjentów nieleczonych bez zmienionych naczyń ($x_1 = 0$), o czym świadczy wspólny wzorzec w przesłance reguły. Inaczej mówiąc, EKG pacjentów z dwiema stenozami leczonych lekiem Z jest równie dobre (stabilne) jak pacjentów bez istotnych zwężeń i bez leczenia. Można w związku z tym wnioskować, że dodatkowe leczenie zileutonem wpłynęło korzystnie na tych pacjentów. Obserwacja ta jest zgodna z hipotezą profesora Andrzeja Szczeklika, według której leki przeciwzapalne mogą wywierać wpływ na czynność elektryczną serca w kontekście teorii zapalnej leżącej u podłoża choroby wieńcowej [151, 143].

W przedstawionym procesie wnioskowania przyjmuje się założenie oparte na medycznej wiedzy dziedzinowej. Wnioski dotyczące korzystnego lub niekorzystnego wpływu czynnika zakłócającego (lek Z) wynikają z przesłanek, według których większe nasilenie choroby wiąże się z większymi zmianami EKG. Im większa liczba istotnych zwężeń tętnic wieńcowych u danego pacjenta, tym bardziej nieprawidłowy zapis jego EKG. Przyjmując takie założenie, pacjenci z dwiema stenozami, mieliby inny (mniej stabilny) wzorzec EKG, gdyby nie otrzymali dodatkowego leczenia i nie znaleźliby się w tej samej grupie, co pacjenci bez stenoz. Reguły krzyżowe nie klasyfikują obiektów, lecz opisują grupy obiektów poddanych odmiennym warunkom. Stanowią zatem sposób prezentacji wiedzy w zrozumiałej formie umożliwiającej interpretację. W przykładzie dotyczącym CNS, mogą być wykorzystane do podjęcia decyzji o kontynuowaniu farmakoterapii w wyznaczonej grupie obiektów.

Jedną z metod generowania reguł decyzyjnych jest tworzenie drzew decyzyjnych. Do budowy *reguł krzyżowych zmian percepcji* zastosowano tzw. *drzewo*

wpływu. Jest to drzewo binarne budowane metodą "góra-dół" algorytmem zachłannym, który dzieli rekurencyjne zbiór obiektów metodą "dziel i rządź". Do podziału danych na dwa podzbiory wybierany jest najlepszy podział w sensie przyjętej miary. Każdy liść *drzewa wpływu* zawiera jedną *regułę krzyżową*, która określa stopień oraz charakter wpływu czynnika zakłócającego na zachowanie i postrzeganie, a zatem klasyfikację obiektów.

8.2.2 Drzewo wpływu

Ogólną zasadę konstrukcji drzew decyzyjnych ukierunkowanych na wzorce opisowe można przedstawić w następujących punktach:

- Zbadanie, czy zbiór obiektów spełnia warunek stopu. Jeśli tak, algorytm kończy pracę, inaczej wykonywana jest dalsza część algorytmu;
- Rozpatrywanie wszystkich możliwych podziałów zbioru obiektów na podzbiory oraz określenie, który z podziałów jest najlepszy na podstawie pewnego, przyjętego kryterium stanowiącego miarę jakości podziału;
- Podział zbioru w najlepszy sposób ze względu na przyjęte kryterium;
- Użycie powyższego algorytmu do wszystkich podzbiorów;
- Zastosowanie drzewa do opisu obiektów.

Miara jakości podziału to funkcja przypisująca podziałowi pewną wartość rzeczywistą, która odzwierciedla jakość podziału w badanym zbiorze obiektów.

W rozprawie jako kryterium wyboru najlepszego podziału zaproponowano miarę opartą na odległości pomiędzy grupami obiektów. Miara ta jest wyliczana z wykorzystaniem dobrze znanego z literatury teorii prawdopodobieństwa pojęcia wartości oczekiwanej zmiennej losowej (ang. *expected value*) [100, 135, 141].

Załóżmy, że zbiór obiektów danej tablicy decyzyjnej $\mathbf{A} = (U, A \cup \{d\})$ zawiera dwie grupy badanych. Jedną z grup poddano działaniu czynnika zakłócającego (grupa Z), a drugą nie (grupa P). Interesuje nas charakterystyka wybranej numerycznej cechy G w obu grupach. W celu jej oceny wykonujemy następujące czynności:

1. Określenie rozkładów prawdopodobieństwa wybranej cechy G w obu grupach, oznaczone jako $G^Z_{\mathbf{A}}$ i $G^P_{\mathbf{A}}$.
2. Zdefiniowanie zmiennej $X_{\mathbf{A}}$ określonej na zbiorach wartości cech G^Z_A i G^P_A stanowiącej różnicę wartości cechy G między grupami. Zbiór wartości cechyX_A jest równy $\{|G^Z_{\mathbf{A}}(o_i) - G^P_{\mathbf{A}}(o_j)|,\ o_i \in Z, o_j \in P\}$.
3. Wyznaczenie rozkładu zmiennej $X_{\mathbf{A}}$.
4. Wyznaczenie wartości oczekiwanej zmiennej $X_{\mathbf{A}}$.

Dla pojęcia choroby niedokrwiennej badaną cechą jest liczba zwężonych naczyń wieńcowych, przyjmująca cztery wartości: $0, 1, 2, 3$ (zależnie od liczby stenoz), a czynnikiem zakłócającym jest dodatkowa terapia lekiem Z. Wartości: $0, 1, 2, 3$ dla każdego pacjenta są reprezentowane w tablicy decyzyjnej $\mathbf{A} = (U, A \cup \{d\})$ przez wartość atrybutu decyzyjnego d.

Rozkłady wartości cechy w obu grupach: poddanej i niepoddanej leczeniu przedstawiają tabele: 8.1 i 8.2.

$G^Z_{\mathbf{A}}$	**0**	**1**	**2**	**3**
$P(G^Z_{\mathbf{A}})$	a_0	a_1	a_2	a_3

Tablica 8.1: Rozkład wartości cechy G w grupie Z czyli poddanej działaniu czynnika zakłócającego.

$G^P_{\mathbf{A}}$	**0**	**1**	**2**	**3**
$P(G^P_{\mathbf{A}})$	b_0	b_1	b_2	b_3

Tablica 8.2: Rozkład wartości cechy G w grupie P czyli bez ekspozycji na czynnik zakłócający.

Rozkład zmiennej $X_{\mathbf{A}}$ (różnica badanej cechy między grupami) przedstawia tabela 8.3, gdzie prawdopodobieństwo p_i, dla $i = 0, ...3$ wyliczane jest następująco:

$$p_0 = P(X_{\mathbf{A}} = 0) = a_0b_0 + a_1b_1 + a_2b_2 + a_3b_3 \tag{8.7}$$
$$p_1 = P(X_{\mathbf{A}} = 1) = a_0b_1 + a_1b_0 + a_1b_2 + a_2b_3 + a_2b_1 + a_3b_2 \tag{8.8}$$
$$p_2 = P(X_{\mathbf{A}} = 2) = a_0b_2 + a_1b_3 + a_2b_0 + a_3b_1 \tag{8.9}$$
$$p_3 = P(X_{\mathbf{A}} = 3) = a_0b_3 + a_3b_0 \tag{8.10}$$

$X_{\mathbf{A}}$	**0**	**1**	**2**	**3**
$P(X_{\mathbf{A}})$	p_0	p_1	p_2	p_3

Tablica 8.3: Rozkład cechy X.

Wartości p_i dla CNS oznaczają prawdopodobieństwo wystąpienia różnicy w liczbie stenoz między grupami równej i. Na przykład, p_2 to prawdopodobieństwo, że pacjenci z lekiem Z mają o 2 stenozy mniej lub więcej od pacjentów bez dodatkowego leczenia.

Wartość oczekiwana zmiennej $X_{\mathbf{A}}$ obliczana według wzoru (8.11) umożliwia ilościowe określenie zróżnicowania wartości badanej cechy w obu grupach i stanowi podstawę do wyznaczenia jakości cięcia w *drzewie wpływu*.

$$E(X_{\mathbf{A}}) = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + 3 \cdot p_3 \tag{8.11}$$

Podczas budowy *drzewa wpływu* poszukujemy takiego podziału, który rozdzieli grupy pacjentów o maksymalnie różnej reakcji na czynnik zakłócający. Do oceny odległości między takimi grupami wyznaczonymi przez cięcie c zaproponowano miarę wyznaczaną według wzoru (8.12):

$$Q_{Impact}(c, \mathbf{A}) = |E(X_{\mathbf{A}(T)}) - E(X_{\mathbf{A}(\neg T)})| \tag{8.12}$$

gdzie $\mathbf{A}(T)$ i $\mathbf{A}(\neg T)$ to podtablice $\mathbf{A}$ zawierające wszystkie obiekty z U pasujące, odpowiednio do wzorca $T = TL(c)$ oraz wzorca $\neg T = TR(c)$. Taka miara zastosowana do budowy drzewa umożliwia ocenę stopnia wpływu czynnika zakłócającego na zachowanie obiektów [33].

Budowa *drzewa wpływu* przebiega według Algorytmu 8.2.1. Podczas działania algorytmu zachłannie wybierane jest cięcie o najwyższej jakości (dla CNS $Q_{Impact} \in [0, 3]$). Jako warunek zakończenia podziałów (warunek stopu) przyjęto przekroczenie wartości oczekiwanej różnicy cechy w badanych grupach $E(X_{\mathbf{A}})$ pewnego zadanego progu t. Dodatkowo podziały należy kończyć także w sytuacji, kiedy liczba obiektów w dzielonym węźle spada poniżej pewnego poziomu. Dla drzewa utworzonego dla danych medycznych przedstawionego na Rys. 8.1 wartość t wynosiła 1.75. W każdym węźle przedstawiono liczbę pacjentów otrzymujących placebo bez istotnych stenoz ($P0$), z jednym istotnym zwężeniem ($P1$), dwoma ($P2$) i trzema zwężeniami ($P3$), jak i tych leczonych zileutonem bez istotnych zwężeń ($Z0$), z jednym ($Z1$), dwoma ($Z2$) i trzema istotnymi zwężeniami ($Z3$). Ponadto przedstawiona jest wartość oczekiwana liczby stenoz (S), osobno w grupie otrzymującej placebo ($E(S|P)$), jak i grupie otrzymującej zileuton ($E(S|Z)$). Dla każdego węzła wyliczono różnicę między tymi wartościami oczekiwanymi (wartość δ) oraz wartość oczekiwaną różnicy w liczbie stenoz między grupą leczoną i nieleczoną ($E(X)$).

Opis: Algorytm rozpoczyna działanie w węźle zawierającym obiekty wejściowej tablicy decyzyjnej. W kroku pierwszym wyznaczany jest najlepszy podział węzła w sensie przyjętej miary jakości podziału zdefiniowanej wzorem (8.12). Cięcie optymalne $c = (a, v)$, wyznaczone w poprzednim kroku, definiuje w węźle dwa wzorce T_c i $\neg T_c$. Do T_c przypisywany jest wzorzec lewy ($TL(c) = \{u \in U : a(u) < v\}$ dla atrybutu numerycznego i $TL(c) = \{u \in U : a(u) = v\}$ dla symbolicznego), natomiast do $\neg T_c$ wzorzec prawy ($TR(c) = \{u \in U : a(u) \geq v\}$ i $TR(c) = \{u \in U : a(u) \neq v\}$ dla odpowiedniego typu atrybutów). Drugi krok polega na podzieleniu obiektów węzła na dwie części, przy czym do pierwszej części trafiają obiekty pasujące do

Algorytm 8.2.1: Konstruowanie drzewa wpływu

WEJŚCIE: Tablica decyzyjna $\mathbf{A} = (U, A \cup \{d\})$, próg jakości podziału węzłów.

WYJŚCIE: *Drzewo wpływu* wyznaczone dla tablicy **A**.

begin

1 Znajdź cięcie optymalne c w tablicy **A** oraz związane z c wzorce $T_c = TL(c)$ i $\neg T_c = TR(c)$. Cięcie optymalne maksymalizuje wartość współczynnika $Q_{Impact}(c, \mathbf{A})$

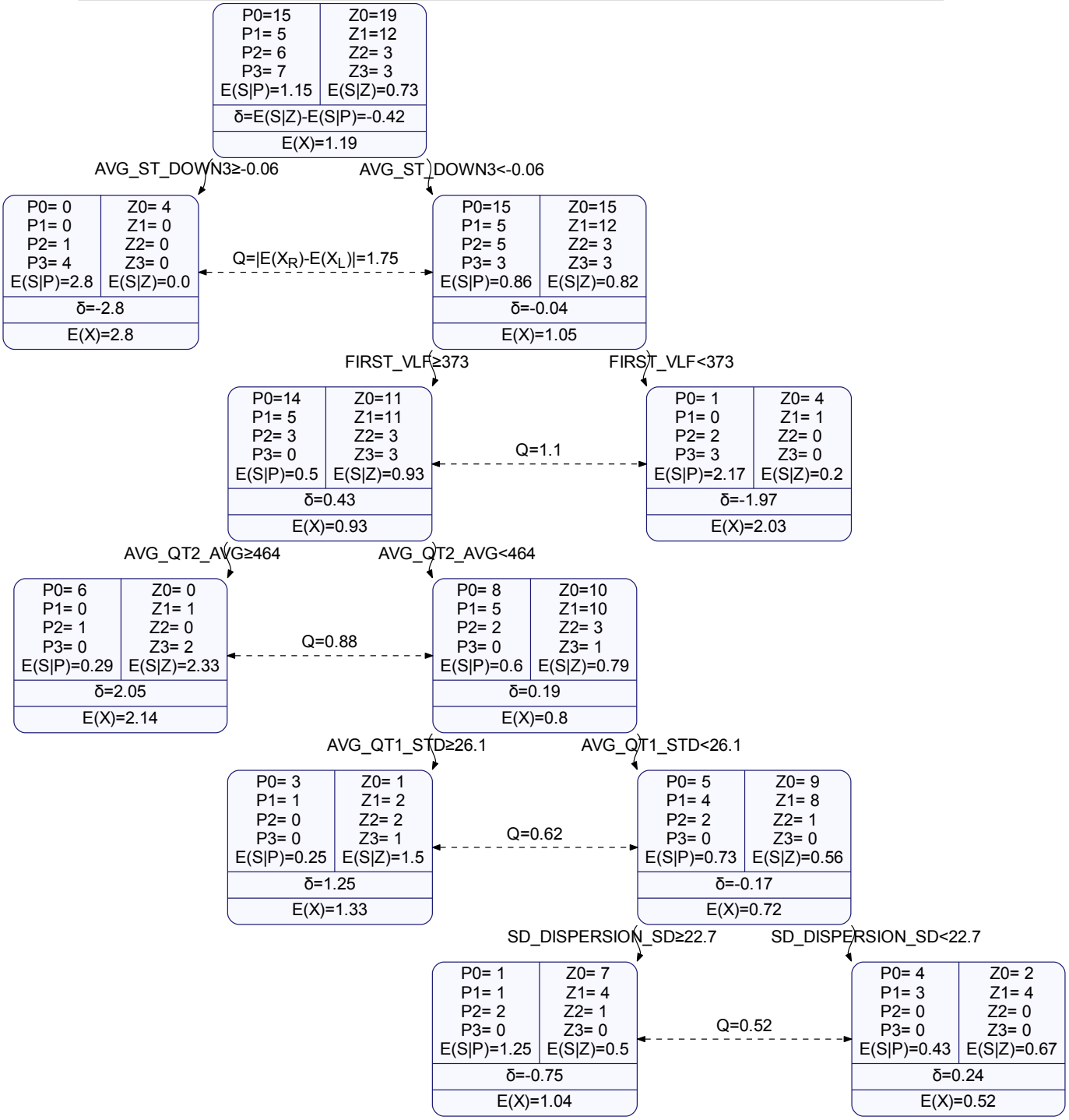
2 Podziel tablicę **A** na dwie podtablice $\mathbf{A}(T_c)$ i $\mathbf{A}(\neg T_c)$ takie, że:
$\mathbf{A}(T_c)$ zawiera obiekty pasujące do wzorca T_c,
$\mathbf{A}(\neg T_c)$ zawiera obiekty pasujące do wzorca $\neg T_c$.

3 **jeżeli** *tablice* $\mathbf{A}(T_c)$ i $\mathbf{A}(\neg T_c)$ *spełniają warunki stopu* **to**
| zakończ tworzenie drzewa
inaczej
| powtarzaj 1-3 dla wszystkich tablic nie spełniających warunku stopu
end

end

wzorca T_c, do prawej zaś - pasujące do $\neg T_c$. W kroku trzecim, węzły są badane pod kątek spełniania warunku zatrzymania podziałów. jeżeli ten warunek jest spełniony, tworzenie drzewa kończy się w danym węźle. Jeżeli natomiast warunek ten nie jest spełniony, wówczas algorytm rozpoczyna pracę od początku, przy czym tablicę wejściową tworzą teraz obiekty przypisane do badanego węzła.

Złożoność obliczeniowa tworzenia *drzewa wpływu* według Algorytmu 8.2.1, w związku z koniecznością sortowania wartości atrybutu wykonywanej w czasie $O(n \cdot \log n)$ dla pojedynczego atrybutu, wynosi: $O(n \cdot m \cdot \log n)$, gdzie n to liczba obiektów, m – liczba atrybutów warunkowych w tablicy **A**.



Rysunek 8.1: Drzewo wpływu utworzone dla danych dotyczących CNS.

8.2.3 Określanie charakteru wpływu czynnika na percepcję

W *drzewie wpływu* interesujące są takie liście (zbiory obiektów), dla których istnieje odpowiednio duże zróżnicowanie pomiędzy grupami obiektów poddanych i niepoddanych działaniu czynnika zakłócającego, wynikające z różnej reakcji na ten czynnik. Wielkość tego zróżnicowania określa w każdym liściu wartość $E(X_{\mathbf{A}})$ obliczana według wzoru (8.11).

Kwestią wrażliwą pozostaje ustalenie interesującego poziomu zróżnicowania grup oraz przypisanie adekwatnego do problemu badawczego opisu (etykiety) dla zachowania obiektów w liściu. Przykładowo dla choroby wieńcowej jako interesującą różnicę uznano wartość $E(X_{\mathbf{A}}) \geq 1.75$ wskazaną przez eksperta dziedzinowego. Oznacza to, że liście, w których pacjenci leczeni lekiem Z mają przeciętnie co najmniej 1.75 stenoz więcej lub mniej niż pacjenci otrzymujący placebo, zawierają obiekty silnie reagujące na dodatkową terapię.

Zmienna $E(X_{\mathbf{A}})$ nie nadaje się jednak do oceny charakteru wpływu czynnika zakłócającego, ponieważ przyjmuje takie same (dodatnie) wartości dla liści z korzystnymi i niekorzystnymi zmianami. Inaczej mówiąc, $E(X_{\mathbf{A}})$ informuje tylko, że w danym liściu występują duże zmiany po ekspozycji na czynnik zakłócający, ale nie wiadomo o jakim charakterze (korzystnym czy niekorzystnym).

W związku z tym do określenia charakteru tego wpływu i przypisania etykiety liściom zaproponowano wyznaczenie wartości δ w danym liściu zgodnie ze wzorem (8.13):

$$\delta = E(dec|Modyfikator = y_2) - E(dec|Modyfikator = y_1) \tag{8.13}$$

Dla problemu choroby niedokrwiennej, sposób działania dodatkowej farmakoterapii przyjmuje więc postać określoną wzorem (8.14):

$$\delta = E(S|terapia = Z) - E(S|terapia = P) \tag{8.14}$$

gdzie $\delta \in [-3, 3]$.
Ogólnie w węzłach *drzewa wpływu* można wyróżnić trzy przypadki:

- $E(dec|Modyfikator = 0) \ll E(dec|Modyfikator = 1)$;
- $E(dec|Modyfikator = 0) \gg E(dec|Modyfikator = 1)$;
- Pozostałe przypadki.

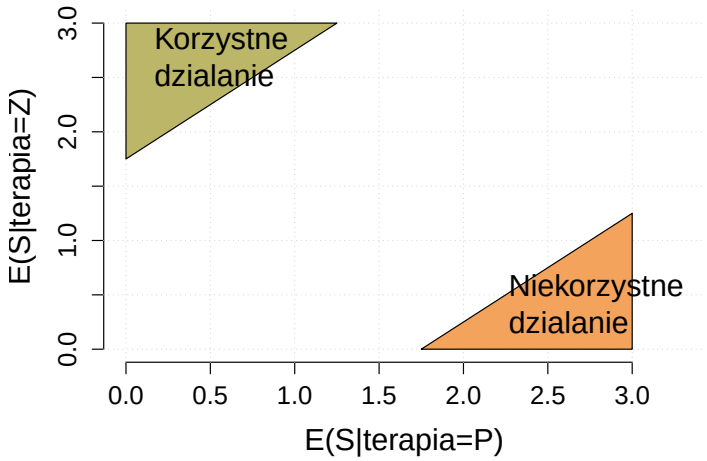
W każdym z powyższych przypadków, wpływ czynnika zakłócającego jest inny, a jego charakter można określić na podstawie wartości δ.

W chorobie niedokrwiennej serca, pierwszy przypadek dotyczy liści, w których δ przyjmuje wartości z przedziału $[1.75, 3]$, przy czym dolna granica tego przedziału została dobrana arbitralnie. Takie wartości δ występują, gdy wartość oczekiwana

liczby zmienionych naczyń w grupie nieleczonej $E(S|terapia = P)$ znajduje się w przedziale $[0, 1.25)$, a $E(S|terapia = Z)$ jest nie mniejsze niż:

$$E(S|terapia = P) + 1.75$$

Odpowiada to zielonemu polu na wykresie wartości oczekiwanej liczby stenoz w grupie leczonej od tej wartości w grupie nieleczonej (Rys. 8.2). W tym przy-



Rysunek 8.2: Zależność wartości oczekiwanej liczby zwężonych naczyń w grupie leczonej od wartości oczekiwanej liczby zwężeń w grupie nieleczonej.

padku można powiedzieć, że czynnik zakłócający spowodował zmianę postrzegania obiektów z istotnie zwężonymi naczyniami, w taki sposób, że są one obserwowane tak jak obiekty z prawidłowymi tętnicami wieńcowymi. Taki liść obejmuje więc przypadki korzystnego wpływu czynnika zakłócającego.

W przypadku drugim, wartość δ znajduje się w przedziale $[-3, -1.75]$, gdzie górną granicę przyjęto arbitralnie. Takie wartości δ występują, gdy wartość oczekiwana liczby zmienionych naczyń w grupie nieleczonej $E(S|terapia = P)$ znajduje się w granicach od 1.75 do 3, a $E(S|terapia = Z)$ jest nie większe niż:

$$E(S|terapia = P) - 1.75$$

Liście o wartości δ z tego przedziału zawierają obiekty, na które czynnik zakłócający zadziałał niekorzystnie. Przykładowo, jeżeli $E(S|terapia = P)$ wynosiłoby 3, a $E(S|terapia = Z)$ 0, to oznaczałoby że pacjenci bez zmienionych tętnic otrzymujący zileuton, mają taki sam wzorzec EKG jak pacjenci z trzema stenozami bez

tego leczenia. Odpowiada to czerwonemu polu na wykresie wartości oczekiwanej liczby stenoz w grupie leczonej od tej wartości w grupie nieleczonej (Rys. 8.2).

W pozostałych przypadkach nie można stwierdzić, czy wpływ czynnika zakłócającego jest korzystny lub niekorzystny.

W rozdziale przestawiono metodę opisu wpływu czynnika zakłócającego percepcję zwężeń w naczyniach krwionośnych. W tym przypadku odkrywanie wiedzy obejmuje wyszukiwanie wzorców charakteryzujących właściwości danych, które są interesujące, użyteczne i zrozumiałe dla użytkownika. Otrzymujemy wzorce opisowe ukierunkowane na interpretację wiedzy pozyskiwanej z danych przez człowieka, przy czym interpretacja ta wymaga udziału wiedzy dziedzinowej.

Możliwości zastosowania tej metody dla przypadku zileutonu obejmują: utrzymywanie chwilowej stabilności w karetce lub w okresie przedoperacyjnym po zawale, czy też stałe podawanie w przypadku udowodnienia skuteczności w konkretnych przypadkach opisanych przez reguły krzyżowe.

Rozdział 9

Badania eksperymentalne

Zawartość

9.1 Charakterystyka danych eksperymentalnych

Główne dane eksperymentalne zostały pozyskane z II Katedry Chorób Wewnętrznych Collegium Medicum Uniwersytetu Jagiellońskiego. Zawierają zapis EKG zarejestrowany metodą Holtera, wzbogacony danymi klinicznymi pacjentów ze sta-

bilną chorobą niedokrwienną serca, z zatokowym rytmem w zapisie EKG. Pacjentów rekrutowano spośrod osób przyjmowanych do Oddziału Chorób Wewnętrznych i Chorób Serca celem wykonania planowego zabiegu koronarografii z ewentualną angioplastyką i implantacją stentu. Bezpośrednio po koronarografii wyniki poddawano analizie angiograficznej, na podstawie której chorych kwalifikowano do leczenia przezskórnego. U chorych zakwalifikowanych do w/w leczenia, wykonywano jednoczasowo angioplastykę wieńcową z lub bez implantacji stentu. Przed i po zabiegu u wszystkich chorych prowadzono 24-godzinne monitorowanie EKG metodą Holtera. Każdorazowo, po zakończeniu badania dane zapisane na przenośnej karcie pamięci zestawu rejestrującego wczytywano do pamięci komputera stacjonarnego, a następnie poddawano automatycznej analizie przy użyciu oprogramowania dostarczonego przez producenta rejestratora EKG. Dane te obejmują dwa zbiory:

- Pierwszy, oznaczony jako $HOLTER\_I$ - zawiera dane 70 pacjentów zebrane w latach 2006-2009. Zapis EKG metodą Holtera przeprowadzono przy użyciu 3-kanałowego zestawu rejestrującego systemu HolCARD 24W firmy Aspel, natomiast dane koronarograficzne zawierają informacje o liczbie istotnie zwężonych tętnic wieńcowych (od 0 do 3). W zbiorze $HOLTER\_I$ wyodrębniono 173 atrybuty warunkowe;

- Drugi - $HOLTER\_II$ - zawiera dane 200 pacjentów zebrane w latach 2015-2016 z wykorzystaniem 12-kanałowego rejestratora R12 systemu BTL CardioPoint-Holter H600 v2-23. Dane angiograficzne zawierają szczegółowe informacje na temat procentowego zwężenia każdej z ocenianych w koronarografii tętnic wieńcowych. Do zbioru wyselekcjonowano pacjentów bez złożonych zaburzeń rytmu serca, takich jak ekstrasystolie nadkomorowe czy komorowe, które uniemożliwiają prawidłową analizę zapisu EKG. Zbiór zawiera 595 atrybutów warunkowych.

Badania związane z rozprawą przeprowadzono na pierwszej części zapisu EKG Holtera (przed koronarografią). Zbiór danych zawierał dokładny opis stanu klinicznego pacjentów (wiek, płeć, rozpoznanie lekarskie), chorób współistniejących, leczenia farmakologicznego, wyniki badań laboratoryjnych (m.in. poziom troponiny I, białka CRP, cholesterolu, LDL) oraz wiele parametrów zapisu holterowskiego dotyczących liczby i rodzaju zaburzeń rytmu, zmian odstępu PQ, zmian odcinka ST czy zmienności rytmu serca HRV (ang. *heart rate variability*) w dziedzinie czasu i częstotliwości oraz zmian odstępu QT. Dane zapisu holterowskiego zostały zagregowane do punktów czasowych opisujących jedną godzinę badania.

Pozyskane dane zostały zapisane w postaci plików binarnych. W pierwszym etapie zweryfikowano kompletność danych pacjentów, następnie utworzono relacyjną bazę danych w środowisku PostgreSQL [116]. Szczegóły na temat implementacji

hurtowni danych znajdują się w Dodatku B. Wczytanie danych do bazy odbyło się przy użyciu importera utworzonego w środowisku Java. W tym procesie dane tekstowe zostały przekształcone do odpowiednich formatów danych określonych osobno dla każdego parametru w taki sposób, aby umożliwić efektywne przechowywanie danych w bazie bez utraty informacji (np. w postaci liczb całkowitych, zmiennoprzecinkowych czy łańcuchów tekstowych). Po wstępnym przetworzeniu danych, takim jak integracja danych poszczególnych pacjentów oraz badań, dane zaimportowano do środowiska Java celem dalszych analiz.

Podstawową charakterystykę oraz dane angiograficzne obydwu zbiorów danych przedstawiają Tab. 9.1 oraz Tab. 9.2.

	HOLTER_I	**HOLTER_II**
Cecha	**Wartość**	**Wartość**
Liczebność (N)	70	200
Wiek	60.6 (38-75)	68.3 (40-89)
Płeć (Mężczyźni/Kobiety)	42 (60%)/28 (40%)	121 (60%)/79 (40%)
Nadciśnienie tętnicze	65 (92.9%)	168 (84%)
Przebyty zawał serca	14 (20%)	79 (39.5%)
Przebyty udar mózgu	2 (2.9%)	brak danych
Miażdżyca tętnic kończyn dolnych	7 (10%)	52 (26%)
Cukrzyca	16 (22.9%)	63 (31.5%)
Palenie papierosów	44 (62.9%)	57 (28.5%)

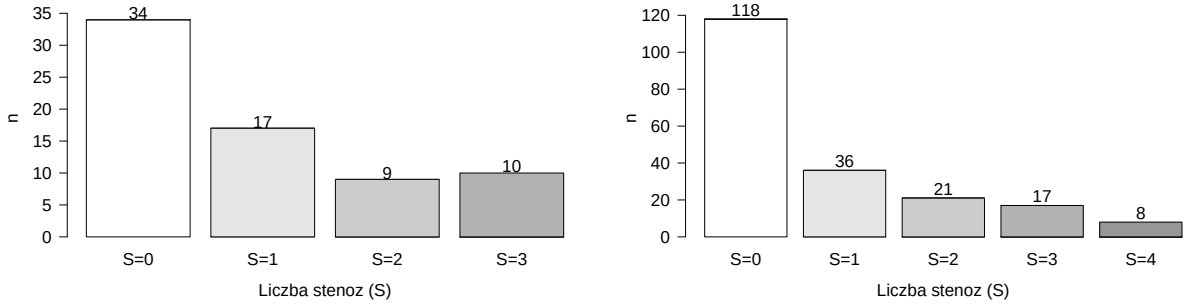
Tablica 9.1: Charakterystyka kliniczna badanych populacji zbioru *HOLTER_I* oraz *HOLTER_II*. Dane przedstawiono jako liczebność (w nawiasach podano %) lub średnią i zakres wartości.

	HOLTER_I	**HOLTER_II**
Wynik koronarografii	N=70 (100%)	N=200 (100%)
Bez istotnych zwężeń w tętnicach wieńcowych	34 (49%)	118 (59%)
Choroba 1-naczyniowa	17 (24%)	36 (18%)
Choroba 2-naczyniowa	9 (13%)	21 (10.5%)
Choroba 3-naczyniowa	10 (14%)	17 (8.5%)
Choroba 4-naczyniowa	-	8 (4%)

Tablica 9.2: Charakterystyka angiograficzna badanych populacji obydwu zbiorów.

Liczbę pacjentów bez istotnych zwężeń oraz z chorobą 1, 2, 3 i 4-naczyniową

w zbiorze *HOLTER_I* oraz *HOLTER_II* przedstawiają odpowiednio Rys. 9.1a i 9.1b.



(a) *HOLTER_I*. (b) *HOLTER_II*.

Rysunek 9.1: Liczba pacjentów bez istotnych zwężeń oraz z chorobą 1, 2, 3 i 4-naczyniową.

Tę samą informację z podziałem na płeć dla obydwu zbiorów zawierają Rys. 9.2a i 9.2b.

(a) *HOLTER_I*. (b) *HOLTER_II*.

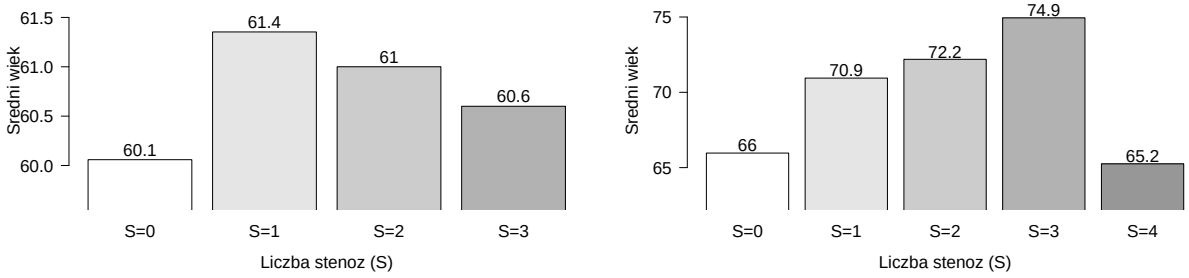
Rysunek 9.2: Liczba pacjentów z podziałem na płeć.

Klasy decyzyjne dotyczące problemu przewidywania obecności istotnych zwężeń tętnic wieńcowych są więc w przybliżeniu równoliczne (Tab. 9.3). Do klasy decyzyjnej "NIE" (brak istotnych zwężeń) należy 49% obiektów zbioru *HOLTER_I*, do klasy "TAK" (obecność istotnych zwężeń) 51% obiektów tego zbioru. Podobnie w zbiorze *HOLTER_II* rozkłady liczności przykładów w klasach są względnie podobne i wynoszą: 59% w klasie decyzyjnej "NIE" oraz 41% w klasie "TAK". Są to zatem dane zrównoważone (zbalansowane).

		HOLTER_I	HOLTER_II
Klasa decyzyjna	**Opis**	N=70 (100%)	N=200 (100%)
Klasa 'NIE'	Brak istotnych zwężeń tętnic wieńcowych	34 (49%)	118 (59%)
Klasa 'TAK'	Obecne istotne zwężenia	36 (51%)	82 (41%)

Tablica 9.3: Rozkład klas decyzyjnych dla problemu predykcji istotnych stenoz tętnic wieńcowych w CNS.

Średni wiek pacjentów dla poszczególnych wyników koronarografii przedstawiono na Rys. 9.3a i 9.3b, natomiast rozkłady wieku pacjentów w chorobie 0, 1, 2, 3 i 4-naczyniowej w badanych zbiorach na Rys. 9.4a oraz 9.4b.

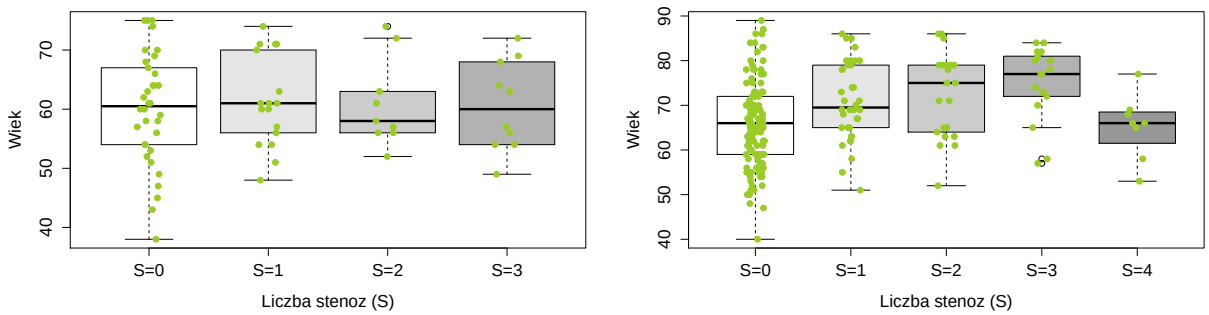


(a) *HOLTER_I*. (b) *HOLTER_II*.

Rysunek 9.3: Średni wiek pacjentów.

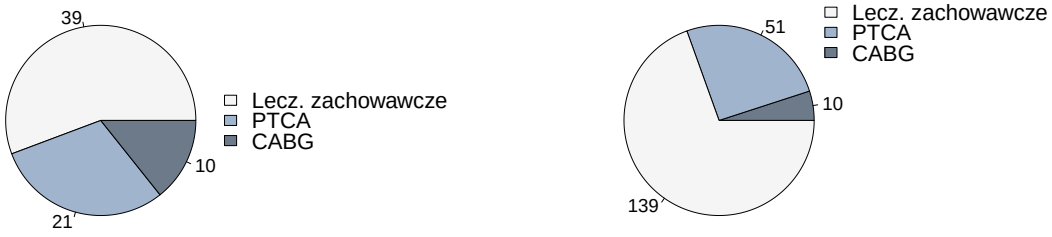
W zbiorze *HOLTER_I* u 30% pacjentów po zabiegu koronarografii wykonano przezskórną angioplastykę PTCA (ang. *percutaneous transluminal coronary angioplasty*), natomiast 14.3% zostało skierowanych do operacji CABG. W pozostałych przypadkach zastosowano leczenie zachowawcze. Spośród pacjentów zbioru *HOLTER_II* zabieg PTCA przeprowadzono u 25.5% z nich, 5% przebyło CABG, natomiast pozostałych 69.5% było leczonych zachowawczo. Ogólną liczbę wykonanych zabiegów w poszczególnych zbiorach przedstawiono na Rys. 9.5a i 9.5b, natomiast odsetki zabiegów dla poszczególnych rodzajów choroby CNS zawierają Rys. 9.6a oraz 9.6b.

53% pacjentów ze zbioru *HOLTER_I* zostało poddanych ponadstandardowej terapii przeciwzapalnej zileutonem (lek Z). Zileuton jako inhibitor 5-lipooksygenazy hamuje biosyntezę leukotrienów, które biorą udział w powstawaniu blaszek miażdżycowych. Podstawową charakterystykę oraz dane angiograficzne pa-



(a) $HOLTER\_I$. (b) $HOLTER\_II$.

Rysunek 9.4: Wiek pacjentów dla różnych rodzajów CNS.



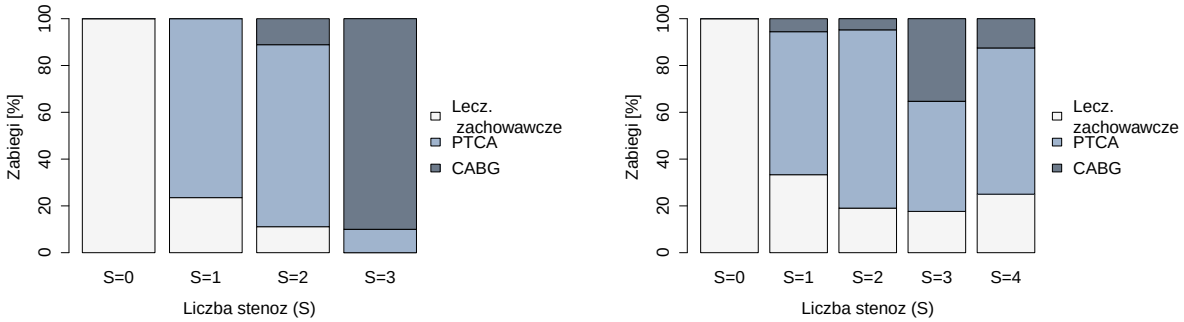
(a) $HOLTER\_I$. (b) $HOLTER\_II$.

Rysunek 9.5: Liczba pacjentów według przeprowadzonego rodzaju leczenia rewaskularyzacyjnego.

cjentów z podziałem na grupę leczoną i nieleczoną lekiem Z przedstawiają Tab. 9.4 oraz Tab. 9.5.

Liczbę pacjentów zbioru $HOLTER\_I$ bez istotnych zwężeń oraz z chorobą 1, 2 i 3-naczyniową w grupie leczonej i nieleczonej lekiem Z przedstawia Rys. 9.7. Natomiast średni wiek pacjentów dla różnych wyników koronarografii w poszczególnych grupach zawiera Rys. 9.8.

Celem eksperymentów była ocena skuteczności metod i algorytmów zaproponowanych w rozprawie w rzeczywistym problemie dotyczącym przewidywania obecności istotnych zwężeń w tętnicach wieńcowych. Do testowania jakości klasyfikatorów w przypadku podzbioru $HOLTER\_I$ osób nieleczonych lekiem Z zastosowano technikę LOO (ang. *leave-one-out*), która zwykle jest stosowana w przypadku małych zbiorów danych oraz k-krotną walidację krzyżową k-CV (ang. *k-fold cross-validation*) dla pozostałych zbiorów. Technika LOO używa jednego obiektu oryginalnego zbioru danych do testowania, a pozostałych obserwacji do trenowa-



(a) *HOLTER_I*. (b) *HOLTER_II*.

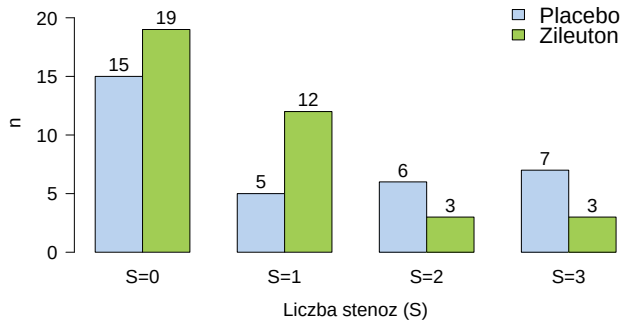
Rysunek 9.6: Odsetek pacjentów według rodzaju leczenia rewaskularyzacyjnego oraz liczby istotnych stenoz.

Cecha	**Placebo**	**Zileuton**
Liczebność (N)	33	37
Wiek	59.4 (38-75)	61.6 (45-75)
Płeć(Mężczyźni/Kobiety)	22/11 (66.7%/ 33.3%)	20/17 (54.1%/ 45.9%)
Nadciśnienie tętnicze	29 (87.9%)	36 (97.3%)
Przebyty zawał serca	4 (12.1%)	10 (27%)
Przebyty udar mózgu	1 (3%)	1 (2.7%)
Miażdżyca tętnic kończyn dolnych	2 (6.1%)	5 (13.5%)
Cukrzyca	7 (21.2%)	9 (24.3%)
Palenie papierosów	21 (63.6%)	23 (62.2%)

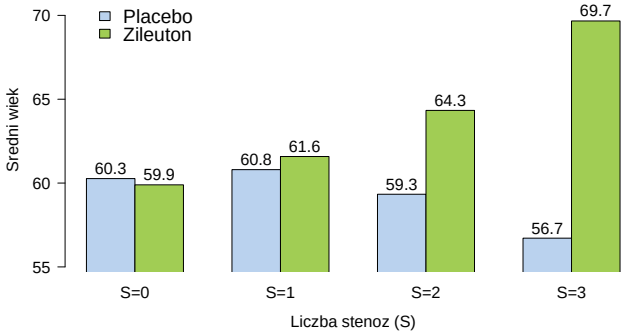
Tablica 9.4: Charakterystyka kliniczna badanych grup ze zbioru *HOLTER_I*. Dane przedstawiono jako liczebność lub średnią (zakres lub frakcję w %).

Wynik koronarografii	**Placebo(N=33)**	**Zileuton(N=37)**
Choroba 0-naczyniowa	15 (45.5%)	19 (51.4%)
Choroba 1-naczyniowa	5 (15.2%)	12 (32.4%)
Choroba 2-naczyniowa	6 (18.2%)	3 (8.1%)
Choroba 3-naczyniowa	7 (21.2%)	3 (8.1%)

Tablica 9.5: Charakterystyka angiograficzna badanych grup ze zbioru *HOLTER_I*.



Rysunek 9.7: Liczba pacjentów zbioru $HOLTER\_I$ bez istotnych zwężeń oraz z chorobą 1, 2 i 3-naczyniową w grupie nieleczonej i leczonej zileutonem.



Rysunek 9.8: Średni wiek pacjentów w grupie nieleczonej i leczonej zileutonem ($HOLTER\_I$).

nia. Taka procedura jest powtarzana n razy w taki sposób, że każda obserwacja w próbce jest wykorzystana jeden raz jako dane testowe. W technice k-CV, zbiór danych jest dzielony na k równych części (ang. *folds*). W każdej spośród k iteracji, k-1 części jest wykorzystanych do uczenia, natomiast pozostała część (za każdym razem inna) do testowania.

Jako miarę sukcesu (lub niepowodzenia) klasyfikacji zastosowano następujące parametry, dobrze znane z literatury: dokładność ACC, pokrycie COV, czułość SN, pokrycie przypadków pozytywnych, precyzję przykładów pozytywnych PPV, specyficzność SP, pokrycie przykładów negatywnych i precyzję przykładów negatywnych NPV. Parametry te opisano szczegółowo w Rozdziale 3.3.

Eksperymenty z metodami I, II i III obejmowały grupę pacjentów ze zbioru

HOLTER_I, którzy nie otrzymywali dodatkowego leczenia zileutonem, stanowiącym czynnik zakłócający zapis EKG (przyczynek do opracowania metody V opisującej modyfikację percepcji) oraz wszystkich ze zbioru *HOLTER_II* (pacjenci nie otrzymywali dodatkowego leczenia zileutonem). W metodach IV i V wykorzystano całe obydwa zbiory danych.

Dodatkowo do testowania *V-drzewa* (metoda III), doświadczenia przeprowadzono na 18 ogólnie dostępnych zbiorach danych pozyskanych z repozytorium *Kent Ridge Biomedical Dataset Repository* [80], repozytorium *UC Irvine Machine Learning Repository* [158] i strony internetowej poświęconej książce *The Elements of Statistical Learning* (Statweb [154]).

Z pierwszego źródła wykorzystano 6 zbiorów, które dotyczą eksperymentów mikromacierzowych wykonanych na materiale biologicznym pochodzącym od pacjentów z guzami jelita grubego (colon tumors [5]), ostrą białaczką limfoblastyczną i szpikową (ALL-AML leukemia [58]), chłoniakiem (lymphoma [4]), rakiem płuc (lung cancer [59]), rakiem jajnika (ovarian cancers [112]) i guzami prostaty (prostate tumors [136]). Tab. 9.6 przedstawia ogólną charakterystykę wykorzystanych danych mikromacierzowych.

Zbiór	**Obiekty**	**Atrybuty**	**Klasy**
lymphoma	47	4026	2
leukemia	72	7129	2
colon	62	2000	2
lung cancer	181	12533	2
prostate	136	12600	2
ovarian cancer	253	15154	2

Tablica 9.6: Charakterystyka danych mikromacierzowych (Kent Ridge Biomedical Dataset Repository).

Kolejnych 12 zbiorów danych pochodzących z repozytorium UCI (dwa pierwsze w Tabeli 9.7) oraz Statweb dotyczy takich zagadnień jak: audiologia, biodegradacja molekuł, sygnały sonaru, pasma na formach drukarskich (cylinder banding), rozpoznawanie chorób rumieniowo-złuszczających (erythemato-squamous diseases), jadalność grzybów, państwa i flagi, detekcja poziomu ozonu, choroba Parkinsona, choroba wieńcowa (SAheart), segmentacja obrazów i spam w poczcie elektronicznej. Tab. 9.7 przedstawia ogólną charakterystykę wykorzystanych zbiorów danych.

Do przeprowadzenia eksperymentów zastosowano własną implementację metod algorytmicznych z biblioteki oprogramowania CommoDM tworzonej w języku Java, jako rozszerzenie biblioteki RS-lib stanowiącej jądro obliczeniowe systemu

Zbiór	Obiekty	Atrybuty	Klasy
audiology	200	71	24
biodeg	1055	41	2
sonar	208	61	2
cylinder_bands	540	40	2
dermatology	366	35	6
mushroom	8124	24	2
flags	194	30	8
ozone	2536	74	2
Parkinson	185	23	2
SAheart	462	9	2
segmentation	2310	20	7
spam	4601	58	2

Tablica 9.7: Charakterystyka eksperymentalnych zbiorów danych (UCI, Statweb).

RSES (jeden z systemów utworzonych w grupie prof. dr. hab. Andrzeja Skowrona z Wydziału Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego) [25].

9.2 Wyniki metody I

9.2.1 Trafność klasyfikacji

Wyniki eksperymentów z klasyfikatorem opartym na drzewie lokalnej dyskretyzacji i wzorcach czasowych zdefiniowanych przez eksperta jako atrybutach warunkowych z dodanymi atrybutami klinicznymi (*CTree-Disc*) do predykcji stenoz wieńcowych w chorobie niedokrwiennej serca (CNS) dla zbioru $HOLTER\_I$ przedstawia Tab. 9.8, natomiast dla zbioru $HOLTER\_II$ Tab. 9.9 (średnie wartości z 10 części procedury 10-CV). Użyte miary jakości zostały zdefiniowane w podrozdziale 3.3, str. 62. Liczbę obiektów prawidłowo i nieprawidłowo sklasyfikowanych w tym eksperymencie przedstawiają Tab. 9.10 oraz Tab. 9.11.

Metoda prawidłowo rozpoznaje 77.8% pacjentów z istotnymi zwężeniami tętnic wieńcowych (czułość, SN) ze zbioru $HOLTER\_I$ oraz 83.4% ze zbioru $HOLTER\_II$. Spośród pacjentów, którzy nie mieli stenoz, metoda prawidłowo identyfikuje 73.3% z nich (specyficzność, SP) w zbiorze $HOLTER\_I$ oraz 93,9% w zbiorze $HOLTER\_II$.

Wartość PPV wynosząca 77.8% dla zbioru $HOLTER\_I$ oraz 88.4% dla $HOLTER\_II$ wskazuje, że wynik pozytywny oznacza wysokie prawdopodobień-

Klasa decyzyjna	Dokładność	Pokrycie	Precyzja
Tak	0.778	1.0	0.778
Nie	0.733	1.0	0.733
Wszystkie klasy (Tak + Nie)	0.758	1.0	-

Tablica 9.8: Wyniki eksperymentów z wykorzystaniem metody I (*CTree-Disc*) do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_I$.

Klasa decyzyjna	Dokładność	Pokrycie	Precyzja
Tak	0.834	0.99	0.884
Nie	0.939	0.99	0.889
Wszystkie klasy (Tak + Nie)	0.894	0.99	-

Tablica 9.9: Wyniki eksperymentów z wykorzystaniem metody I (*CTree-Disc*) do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_II$.

		Sklasyfikowane	
		TAK	NIE
Rzeczywiste	TAK	14	4
	NIE	4	11

Tablica 9.10: Macierz pomyłek klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_I$.

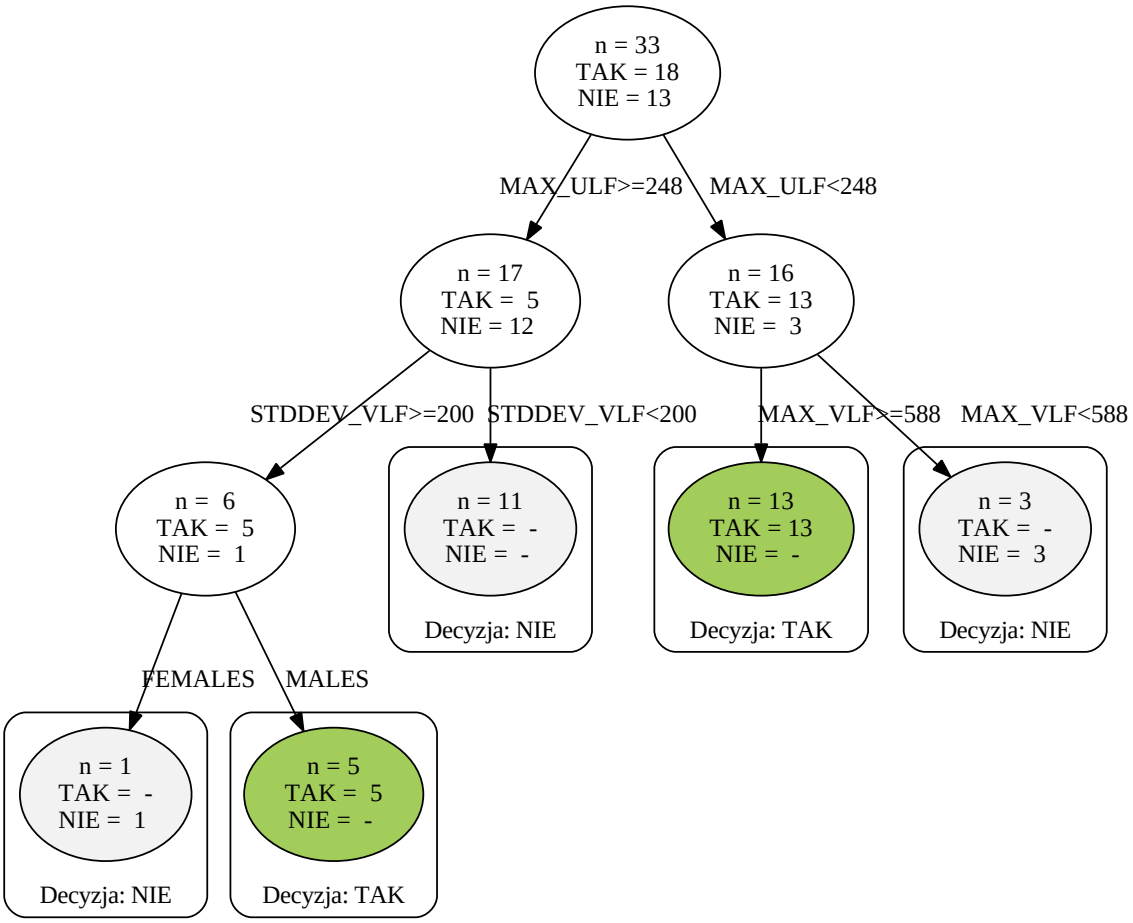
		Sklasyfikowane	
		TAK	NIE
Rzeczywiste	TAK	68	13
	NIE	8	109

Tablica 9.11: Macierz pomyłek klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_II$.

stwo obecności choroby, a test jest dobry w potwierdzeniu zwężenia tętnic wieńcowych. Wynik ujemny w 73.3% w przypadku zbioru $HOLTER\_I$ oraz 88.9% dla $HOLTER\_II$ potwierdza, że pacjent nie ma stenoz wieńcowych (NPV).

W trakcie zastosowanej procedury LOO do oceny jakości klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_I$, większość generowanych drzew wykazywała tę

samą topologię jak ostateczne drzewo decyzyjne (wygenerowane dla całego zbioru) przedstawione na Rys. 9.9, zachowując kolejność potomków i przodków poszczególnych węzłów. Topologia reszty drzew była podobna, to znaczy, były pewne różnice w przypadku wartości atrybutów w węzłach drzewa, a czasami w atrybutach na niższych poziomach generowanych drzew. Na rysunku w każdym węźle drzewa podano liczbę obiektów tworzących węzeł oraz rozkład klas decyzyjnych.



Rysunek 9.9: Drzewo decyzyjne otrzymane metodą I do predykcji stenoz w CNS dla zbioru $HOLTER\_I$.

Przykładowo, 85% drzew do najlepszego cięcia wyznaczającego podział obiektów w korzeniu drzewa wybierało atrybut MAX_ULF (maksymalna w oknie moc pasma o najniższej częstotliwości ULF w widmie zmienności rytmu serca HRV), a 79% drzew za najlepsze cięcie wybierało parę (MAX_ULF, 248). Z kolei najlepsze cięcie, wyznaczające podział lewego poddrzewa korzenia, w przypadku 76% drzew zdefiniowane było na bazie atrybutu $STDDEV\_VLF$, a podział prawego

poddrzewa na bazie atrybutu MAX_VLF w 97% drzew.

W podejmowaniu decyzji najważniejsze były więc wartości mocy widma zmienności rytmu serca HRV w paśmie o bardzo niskiej (VLF) i najniższej częstotliwości (ULF) oraz płeć pacjenta (94% drzew). Co ciekawe, dla tych parametrów analizy widmowej HRV brak jasno zdefiniowanych wartości granicznych normy. O tych miarach wiadomo natomiast, że wykazują istotny związek ze śmiertelnością po zawałach mięśnia sercowego [26]. Wartości tych parametrów wskazane przez zaproponowaną metodę mogą dostarczać pewnych wskazówek przy wyznaczaniu wartości granicznych.

Pomiary analizy widmowej zmienności rytmu serca są szczególnie atrakcyjnymi kandydatami do predykcji śmiertelności, ponieważ pewne pasma częstotliwości są związane z kontrolą rytmu węzła zatokowego (naturalny rozrusznik serca) przez autonomiczny układ nerwowy. Dlatego też uznaje się, że miary te mają potencjał, aby zapewnić wgląd w mechanizmy śmierci, jak również do przewidywania śmiertelności. Osłabienie zmienności rytmu serca po zawale może być związane ze zmniejszeniem aktywności nerwu błędnego w sercu, co prowadzi do przewagi układu współczulnego i w konsekwencji do niestabilności elektrycznej serca. Innym wyjaśnieniem jest zmniejszona reakcja węzła zatokowego na bodźce nerwowe [153].

W przypadku zbioru $HOLTER\_II$, podczas procedury 10-CV, wszystkie drzewa dzieliły korzeń z użyciem cięcia: $(mi, 0.5)$, czyli na podstawie przebytego wcześniej zawału mięśnia sercowego (łac. *myocardial infarctus*), przy czym istotne było rozróżnienie między brakiem zawału i jego przebyciem, bez względu na liczbę przebytych zawałów mięśnia sercowego (1 lub 2). Selekcja dokonana przez model jest więc tutaj zgodna z wiedzą dziedzinową. Duże znaczenie tej cechy pokrywa się z uznawaniem jej za jeden z czynników prognostycznych nagłego zgonu sercowego, czyli niekorzystnego przebiegu np. choroby serca. Do atrybutów, wybieranych często na wyższych poziomach drzew podczas CV należały też: płeć, parametry analizy HRV w dziedzinie czasu, takie jak SDNN (ang. *standard deviation of NN*), czyli odchylenie standardowe czasów trwania wszystkich odstępów NN w badanym okresie, gdzie NN (ang. *normal–normal*) to odstęp pomiędzy kolejnymi prawidłowymi pobudzeniami zatokowymi oraz pomiary odstępu PQ. Zmiany odstępu PQ mogą występować w zaburzeniach rytmu serca. W drzewie zbudowanym dla całego zbioru $HOLTER\_II$, większość podziałów oparta była na tych samych atrybutach, a wiele także na tych samych wartościach, jak dla drzew tworzonych podczas walidacji krzyżowej. Pokazuje to, że metoda ta jest stosunkowo odporna na szum w danych.

Porównano także efektywność metody I (*CTree-Disc*) z innymi powszechnie znanymi metodami klasyfikacji, takimi jak naiwny klasyfikator Bayesa NB (ang. *naive Bayes*), drzewa DT i reguły decyzyjne DR (ang. *decision trees, decision rules*), metoda wektorów nośnych SVM (ang. *supported vector machines*), metoda k

najbliższych sąsiadów k-NN, sztuczne sieci neuronowe ANN (ang. *artifitial neural networks*) oraz lasy losowe RF (ang. *random forests*). Wykorzystano implementację tych metod z następujących systemów znanych dobrze z literatury: WEKA [156], RSES [25, 128], ROSE2 [155] (zastosowano wczesną implementację algorytmu ModLEM [99] dostępną w ROSE2). Należy zaznaczyć, że wyniki pochodzące z systemów WEKA oraz ROSE2 były wygenerowane za pomocą standardowych ustawień. Wyniki eksperymentów dla zbioru $HOLTER\_I$ przedstawia Tab. 9.26. Pokrycie wszystkich przedstawionych metod dla tego zbioru wynosiło 1.0 (każdy obiekt został sklasyfikowany). Spośród testowanych metod (bez dodatkowej wiedzy dziedzinowej WD) jedynie SN metody k-NN była lepsza niż dla proponowanej w rozprawie metody *CTree-Disc*. Wartości ACC, SN, SP, PPV czy NPV wszystkich innych testowanych metod nie przekroczyły wartości tych wskaźników uzyskanych dla metody *CTree-Disc*. Wyniki eksperymentów dla zbioru $HOLTER\_II$ przedstawia natomiast Tab. 9.27. Brak w tej tabeli wartości dla metody "global discretiztion + all ruls (RSES)" wynika z ograniczenia pamięci operacyjnej potrzebnej do wykonania tej operacji na zbiorze $HOLTER\_II$. Spośród porównywanych metod dla tego zbioru danych tylko metoda ModLEM uzyskała lepszą od metody I dokładność i precyzję klasyfikacji, a C4.5 i SVM - SN i NPV. We wszystkich pozostałych przypadkach metoda II była lepsza od metod bez dodatkowej wiedzy dziedzinowej WD.

9.2.2 Analiza statystyczna wyników

Sposób gromadzenia danych wykorzystanych w eksperymencie świadczy o tym, że można je uznać za losowe z całej populacji pacjentów trafiających do kliniki w okresie wieloletnim. Przeprowadzono weryfikację hipotezy o braku skorelowania w całej populacji faktycznych wartości decyzji i wartości decyzji proponowanych przez metodę I (H_0: korelacja=0). Próba liczy 33 osoby a przyjmuje się, że test χ^2 można stosować dla prób >=20. Na podstawie macierzy pomyłek zbudowano macierz liczności rzeczywistych i liczności teoretycznych. Odpowiednie macierze dla zbioru $HOLTER\_I$ przedstawiają Tab. 9.12 oraz 9.13, gdzie $B_{ij} = (\sum A_i \cdot \sum A_j)/\sum A$. Testy statystyczne przeprowadzono za pomocą oprogramowania Statistica StatSoft [146]. Ponieważ wartości teoretyczne są mniejsze od 10, dlatego w teście istotności korelacji uwzględniono poprawkę Yatesa. Wg programu Statistica progowa wartość współczynnika istotności, przy której hipotezę o braku skorelowania badanych cech należy odrzucić, wynosi p=0.0097. Tak więc dla dowolnego poziomu ufności < 0.9983 należy odrzucić hipotezę o braku skorelowania faktycznych wartości decyzji i wartości decyzji proponowanych przez klasyfikator.

Natomiast macierze liczności rzeczywistych i liczności teoretycznych dla zbioru $HOLTER\_II$ przedstawiają Tab. 9.14 i 9.15. Ponieważ część wartości teoretycz-

A	TAK	NIE	Suma
TAK	14	4	18
NIE	4	11	15
Suma	18	15	33

Tablica 9.12: Macierz liczności rzeczywistych wyników klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_I$.

B	TAK	NIE
TAK	$(18 \cdot 18)/33 = 9.82$	$(18 \cdot 15)/33 = 8.18$
NIE	$(15 \cdot 18)/33 = 8.18$	$(15 \cdot 15)/33 = 6.82$

Tablica 9.13: Macierz liczności teoretycznych wyników klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_I$.

A	TAK	NIE	Brak klasyfikacji	Suma
TAK	68	13	1	82
NIE	8	109	1	118
Suma	76	122	2	200

Tablica 9.14: Macierz liczności rzeczywistych wyników klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_II$.

B	TAK	NIE	Brak klasyfikacji
TAK	31.16	50.02	0.82
NIE	44.84	71.98	1.18

Tablica 9.15: Macierz liczności teoretycznych wyników klasyfikatora *CTree-Disc* dla zbioru $HOLTER\_II$.

nych jest < 10 więc należy stosować test χ^2 z poprawką. Jednak poprawka Yatesa może być stosowana tylko do tablic 2x2. Dlatego (na potrzeby prowadzonej weryfikacji) obiekty niesklasyfikowane zostały uznane jako błędnie sklasyfikowane. Dopiero do tak zmodyfikowanej tablicy liczności rzeczywistych zastosowano test χ^2. Wg programu Statistica progowa wartość współczynnika istotności, przy której

hipotezę o braku skorelowania badanych cech w całej populacji należy odrzucić, jest mniejsza od p=0.0001. Tak więc dla dowolnego poziomu ufności <= 0.9999 należy odrzucić hipotezę o braku skorelowania faktycznych wartości decyzji i wartości decyzji proponowanych przez klasyfikator.

9.3 Wyniki metody II

9.3.1 Trafność klasyfikacji

Wyniki eksperymentów z klasyfikatorem opartym na drzewie lokalnej dyskretyzacji ze zmodyfikowaną miarą jakości podziałów (*CTree-DiscW*) do predykcji stenoz wieńcowych w chorobie niedokrwiennej serca dla danych ze zbioru $HOLTER\_I$ przedstawia Tab. 9.16, natomiast dla zbioru $HOLTER\_II$ - Tab. 9.17.

Klasa decyzyjna	**Dokładność**	**Pokrycie**	**Precyzja**
Tak	0.944	1.0	0.85
Nie	0.8	1.0	0.923
Wszystkie klasy (Tak + Nie)	0.879	1.0	-

Tablica 9.16: Wyniki eksperymentów z wykorzystaniem metody II (*CTree-DiscW*) do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_I$.

Klasa decyzyjna	**Dokładność**	**Pokrycie**	**Precyzja**
Tak	0.856	0.99	0.875
Nie	0.924	0.99	0.89
Wszystkie klasy (Tak + Nie)	0.883	0.99	-

Tablica 9.17: Wyniki eksperymentów z wykorzystaniem metody II (*CTree-DiscW*) do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_II$.

Liczbę obiektów prawidłowo i nieprawidłowo sklasyfikowanych w tym eksperymencie przedstawiają Tab. 9.18 oraz Tab. 9.19.

Dla danych ze zbioru $HOLTER\_I$ metoda *CTree-DiscW* prawidłowo identyfikuje 94.4% pacjentów ze stenozą (SN) oraz 80% spośród tych, u których nie było istotnych zwężeń tętnic wieńcowych (SP). W porównaniu do metody I, dokładność i precyzja klasyfikacji wzrosły między 9% a 21%. Zwraca uwagę SN wynosząca

		Sklasyfikowane	
		TAK	NIE
Rzeczywiste	TAK	17	1
	NIE	3	17

Tablica 9.18: Macierz pomyłek klasyfikatora *CTree-DiscW* dla zbioru $HOLTER\_I$.

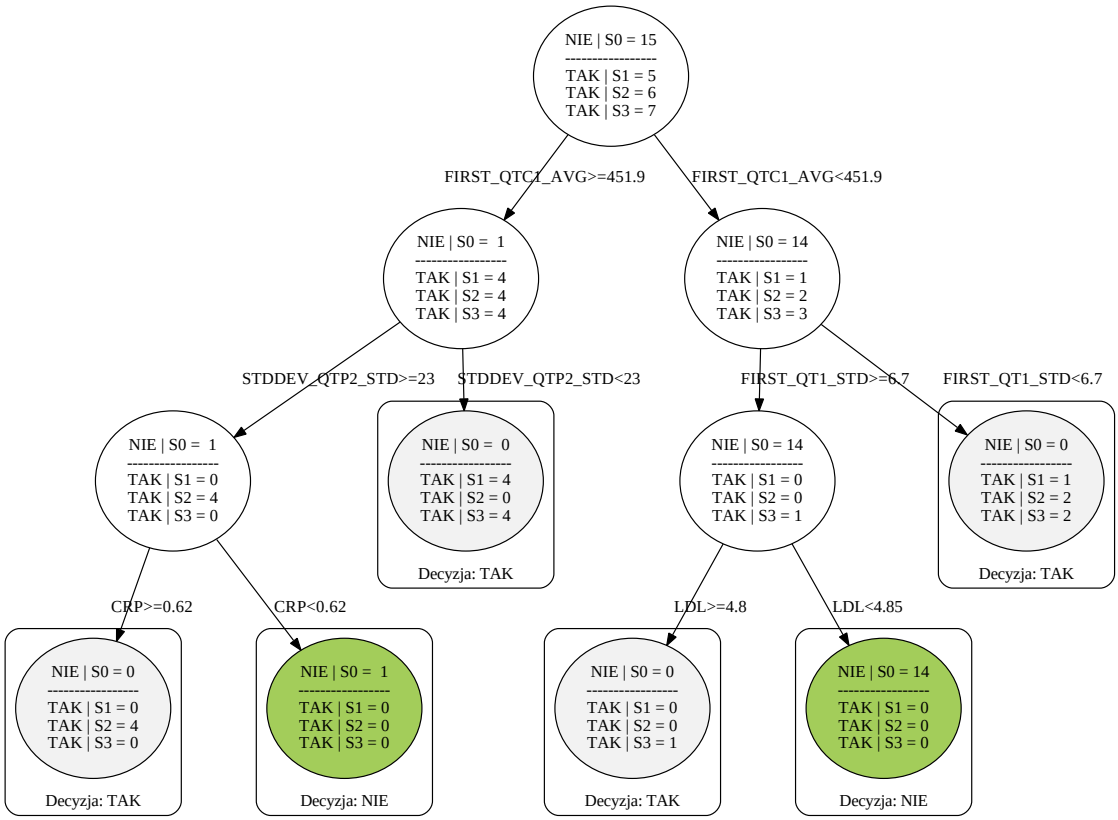
		Sklasyfikowane	
		TAK	NIE
Rzeczywiste	TAK	68	13
	NIE	10	107

Tablica 9.19: Macierz pomyłek klasyfikatora *CTree-DiscW* dla zbioru $HOLTER\_II$.

94.4%, co jest szczególnie istotne w diagnostyce medycznej. Wartość PPV wynosząca 85% wskazuje, że pozytywny wynik testu z dużym prawdopodobieństwem potwierdza stenozę, natomiast wynik negatywny klasyfikacji z jeszcze większym prawdopodobieństwem potwierdza brak zwężeń naczyń wieńcowych (NPV=92.3%). W podejmowaniu ostatecznej decyzji najważniejsze były: czas trwania odstępów QT w zapisie Holtera, poziom CRP (białko C-reaktywne) oraz LDL (frakcja lipoprotein o niskiej gęstości) w surowicy. Ostateczne drzewo decyzyjne przedstawiono na Rys. 9.10. W każdym węźle drzewa podano liczbę obiektów z poszczególnych klas decyzyjnych z uwzględnieniem informacji na temat wewnętrznego zróżnicowania klas, gdzie $S0$ oznacza liczbę obiektów bez zmienionych naczyń (liczba stenoz S równa 0), $S1$ liczbę obiektów z jedną stenozą, $S2$ z dwoma zwężeniami, a $S3$ liczbę obiektów węzła z trzema stenozami.

W przypadku zbioru $HOLTER\_II$ uzyskano prawidłową identyfikację 85.6% chorych z istotnymi stenozami oraz 92.4% pacjentów bez takich zwężeń naczyń wieńcowych. Podobnie do metody I dla tego zbioru wyselekcjonowane zostały takie atrybuty jak: przebyty zawał mięśnia sercowego, parametry analizy HRV, płeć i CRP.

Porównanie efektywności metody II (*CTree-DiscW*) z innymi powszechnie znanymi metodami klasyfikacji (naiwny klasyfikator Bayesa NB, drzewa DT i reguły decyzyjne DR, metoda wektorów nośnych SVM, metoda k-NN, sztuczne sieci neuronowe ANN oraz lasy losowe RF) dla zbioru $HOLTER\_I$ przedstawia Tab. 9.26.



Rysunek 9.10: Drzewo decyzyjne otrzymane metodą II do predykcji stenoz w CNS dla zbioru $HOLTER\_I$.

Wartości wszystkich wyznaczonych miar dokładności klasyfikacji (ACC, SN, SP, PPV i NPV) testowanych metod (bez dodatkowej wiedzy dziedzinowej WD) nie osiągnęły wartości tych wskaźników otrzymanych dla metody *CTree-DiscW*. Dla zbioru $HOLTER\_II$ (Tab. 9.27) spośród porównywanych metod tylko ModLEM uzyskała lepszą od metody I, dokładność (ACC) i specyficzność (SP) klasyfikacji.

9.3.2 Analiza statystyczna wyników

Podobnie jak w przypadku metody pierwszej przeprowadzono weryfikację hipotezy o braku skorelowania faktycznych wartości decyzji i wartości decyzji proponowanych przez metodę II (H_0: korelacja=0). Na podstawie macierzy pomyłek zbudowano macierz liczności rzeczywistych i liczności teoretycznych. Odpowiednie macierze dla zbioru $HOLTER\_I$ przedstawiają Tab. 9.20 oraz 9.21, gdzie $B_{ij} = (\sum A_i \cdot \sum A_j)/\sum A$. Ponieważ wartości teoretyczne są mniejsze od 10, dlatego w teście istotności korelacji zastosowano poprawkę Yatesa. Wg programu

A	**TAK**	**NIE**	**Suma**
TAK	17	1	18
NIE	3	12	15
Suma	20	13	33

Tablica 9.20: Macierz liczności rzeczywistych wyników klasyfikatora *CTree-DiscW* dla zbioru $HOLTER\_I$.

B	**TAK**	**NIE**
TAK	$(18 \cdot 20)/33 = 10.91$	$(18 \cdot 13)/33 = 7.09$
NIE	$(15 \cdot 20)/33 = 9.09$	$(15 \cdot 13)/33 = 5.91$

Tablica 9.21: Macierz liczności teoretycznych wyników klasyfikatora *CTree-DiscW* dla zbioru $HOLTER\_I$.

Statistica progowa wartość współczynnika istotności, przy której hipotezę o braku skorelowania badanych cech należy odrzucić, wynosi p=0.0001. Tak więc dla dowolnego poziomu ufności < 0.9999 należy odrzucić hipotezę o braku skorelowania faktycznych wartości decyzji i wartości decyzji proponowanych przez klasyfikator.

Macierze liczności rzeczywistych i teoretycznych dla zbioru $HOLTER\_II$ przedstawiają Tab. 9.22 oraz 9.23. Dalszy ciąg analizy jest analogiczny jak opisano

A	**TAK**	**NIE**	**Brak klasyfikacji**	**Suma**
TAK	68	13	1	82
NIE	10	107	1	118
Suma	78	120	2	200

Tablica 9.22: Macierz liczności rzeczywistych wyników klasyfikatora *CTree-DiscW* dla zbioru $HOLTER\_II$.

w podrozdziale 9.2.2 przy weryfikacji metody I dla danych $HOLTER\_II$. Macierze pomyłek 9.11 i 9.19 są na tyle nieznacznie różne, że spodziewany wniosek tej analizy jest identyczny z otrzymanym w podrozdziale 9.2.2. Przeprowadzona w programie Statistica weryfikacja hipotezy o braku skorelowania pomiędzy faktycznymi wartościami decyzji a wartościami proponowanymi przez klasyfikator potwierdza, że dla dowolnego poziomu ufności < 0.999 należy odrzucić tę hipotezę.

B	**TAK**	**NIE**	**Brak klasyfikacji**
TAK	31.98	49.2	0.82
NIE	46.02	70.8	1.18

Tablica 9.23: Macierz liczności teoretycznych wyników klasyfikatora *CTree-DiscW* dla zbioru $HOLTER\_II$.

9.4 Wyniki metody III

9.4.1 Trafność klasyfikacji

Celem kolejnych eksperymentów było sprawdzenie jakości klasyfikacji dokonanej przez *V-drzewa decyzyjne*, czyli drzewa z cięciami weryfikującymi. Wyniki doświadczeń przeprowadzonych z wykorzystaniem drzewa *VTree-Disc* do predykcji liczby istotnych zwężeń tętnic wieńcowych w CNS dla zbioru $HOLTER\_I$ przedstawia Tab. 9.24. Przy wysokiej czułości metody III wynoszącej 94.4%, uzyskano poprawę SP o 8% w odniesieniu do metody II dla tego zbioru.

Klasa decyzyjna	**Dokładność**	**Pokrycie**	**Precyzja**
Tak	0.944	1.0	0.894
Nie	0.866	1.0	0.928
Wszystkie klasy (Tak + Nie)	0.909	1.0	-

Tablica 9.24: Wyniki eksperymentów z wykorzystaniem metody III (*VTree-Disc*) do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_I$.

Wyniki *V-drzewa* dla zbioru $HOLTER\_II$ zawiera Tab. 9.25.

Klasa decyzyjna	**Dokładność**	**Pokrycie**	**Precyzja**
Tak	0.902	1.0	0.925
Nie	0.949	1.0	0.933
Wszystkie klasy (Tak + Nie)	0.930	1.0	-

Tablica 9.25: Wyniki eksperymentów z wykorzystaniem metody III (*VTree-Disc*) do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_II$.

Eksperymenty przeprowadzono z następującymi wartościami parametrów:

- Liczba weryfikujących cięć k=5;
- Minimalna jakość cięć weryfikujących: 0.9;
- Metoda wyszukiwania najlepszych cięć: liczba rozróżnianych par obiektów (DiscPairs).

Tak jak w przypadku metody I i II, porównano efektywność metody III z innymi metodami klasyfikacji (naiwny klasyfikator Bayesa NB, drzewa DT i reguły decyzyjne DR, metoda wektorów nośnych SVM, metoda k najbliższych sąsiadów

k-NN, sztuczne sieci neuronowe ANN oraz lasy losowe RF) z wykorzystaniem danych medycznych. Wyniki eksperymentów dla zbioru $HOLTER\_I$ przedstawia Tab. 9.26. Dokładność ACC, specyficzność SP, PPV czy NPV żadnej z innych

Metoda	Dokładność			Precyzja	
	Wszystkie klasy	Tak	Nie	Tak	Nie
C4.5 (WEKA)	0.545	0.611	0.467	0.579	0.500
NaiveBayes (WEKA)	0.394	0.611	0.133	0.458	0.222
SVM (WEKA)	0.545	0.611	0.467	0.579	0.500
k-NN (WEKA)	0.667	**0.833**	0.467	0.652	0.700
RandomForest (WEKA)	0.515	0.722	0.267	0.542	0.444
Multilayer Perceptron (WEKA)	0.548	0.611	0.467	0.579	0.500
Global discretization + all rules (RSES)	0.667	0.611	0.733	0.733	0.611
Local discretization + all rules (RSES)	0.758	0.778	0.733	0.778	0.733
ModLEM (ROSE2)	0.576	0.556	0.600	0.625	0.529
CTree-Disc	**0.758**	**0.778**	**0.733**	**0.778**	**0.733**
CTree-DiscW	**0.879**	**0.944**	**0.8**	**0.85**	**0.923**
VTree-Disc	**0.909**	**0.944**	**0.866**	**0.894**	**0.928**

Tablica 9.26: Wyniki porównawcze z wykorzystaniem innych metod klasyfikacji do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_I$.

testowanych metod (bez dodatkowej wiedzy dziedzinowej WD) nie przekroczyła wartości tych wskaźników dla proponowanych w rozprawie metod: I, II ani III. Najwyższą dokładność ze wszystkich metod (z lub bez dodatkowej WD) uzyskała metoda III, a więc *V-drzewa* decyzyjne z cięciami weryfikującymi.

Wyniki eksperymentów dla zbioru $HOLTER\_II$ przedstawia natomiast Tab. 9.27. Żadna z innych testowanych metod (bez WD) nie osiągnęła dokładności najlepszej z proponowanych metod, tj. *V-drzewa*. Metody: C4.5, SVM oraz ModLEM osiągnęły dokładność zbliżoną do wyników proponowanych metod, natomiast tylko metoda ModLEM (90.5%) osiągnęła dokładność na poziomie średniej dokładności metod: I, II i III (90.2%). Dokładność metod: I, II i III jest dla tego zbioru danych średnio większa o 14% w odniesieniu do średniej dokładności innych metod.

Dodatkowo do testowania *V-drzewa* wykorzystano 18 ogólnodostępnych zbiorów danych powszechnie stosowanych w eksploracji danych. W eksperymentach tych zastosowano 10-krotną walidację krzyżową (10-CV) z powtórzeniem całej procedury 10 razy.

Metoda	Dokładność			Precyzja	
	Wszystkie klasy	Tak	Nie	Tak	Nie
C4.5 (WEKA)	0.875	**0.841**	0.898	0.852	**0.891**
NaiveBayes (WEKA)	0.560	0.329	0.720	0.450	0.607
SVM (WEKA)	0.860	**0.854**	0.864	0.814	**0.895**
k-NN (WEKA)	0.665	0.500	0.780	0.612	0.692
RandomForest (WEKA)	0.750	0.561	0.881	0.767	0.743
Multilayer Perceptron (WEKA)	0.825	0.805	0.839	0.776	0.861
Global discretization + all rules (RSES)	-	-	-	-	-
Local discretization + all rules (RSES)	**0.885**	**0.919**	0.817	**0.902**	0.864
ModLEM (ROSE2)	**0.905**	0.825	**0.960**	**0.932**	**0.890**
CTree-Disc	**0.894**	**0.834**	**0.939**	**0.884**	**0.889**
CTree-DiscW	**0.883**	**0.856**	**0.924**	**0.875**	**0.89**
VTree-Disc	**0.930**	**0.902**	**0.949**	**0.925**	**0.933**

Tablica 9.27: Wyniki porównawcze z wykorzystaniem innych metod klasyfikacji do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_II$.

W eksperymentach wykorzystano 3 miary jakości podziałów węzłów drzewa opisane w Rozdziale 6.2 oparte na: liczbie rozróżnianych par obiektów *DiscPairs*, entropii i indeksie Giniego oraz prostą metodę rozwiązywania konfliktów między cięciami, tj. głosowanie większościowe. Dla każdej z trzech miar porównano jakość klasyfikacji *V-drzewa* (nazywanego dla poszczególnych miar: QV_{Disc}, $QV_{Entropy}$, QV_{Gini} odpowiednio klasyfikatorem: *VTree-Disc*, *VTree-Entropy*, *VTree-Gini*) z jakością klasyfikatora opartego na drzewie lokalnej dyskretyzacji, nazwanego drzewem klasycznym (*CTree*), w szczególności *CTree-Disc*, *CTree-Entropy*, *CTree-Gini* z miarami opisanymi w Rozdziale 3.1.2, tj. Q_{Disc}, $Q_{Entropy}$ oraz Q_{Gini} odpowiednio.

Ostateczne wyniki klasyfikacji dla wybranych miar jakości cięć zawarte w Tab. 9.28, 9.29 i 9.30 stanowią średnią wartość z 10 powtórzeń. W związku z obecnością wartości brakujących w niektórych zbiorach danych, wartość pokrycia jest mniejsza od 1 w kilku przypadkach.

Rysunek 9.11 przedstawia porównanie wyników uzyskanych przez każdą z miar w drzewach. Jak widać, miara *DiscPairs* (oznaczona jako *P*) i *Entropia* (oznaczona jako *E*) uzyskały najlepsze wyniki w przypadku ośmiu zbiorów danych w klasycznym podejściu, podczas gdy w *V-drzewie* zdecydowanie najlepsze wyniki osiągnął algorytm wykorzystujący miarę opartą na liczbie rozróżnianych par z różnych klas

Zbiór	Klasyfikator *CTree-Disc*				Klasyfikator *VTree-Disc*			
	ACC	SD	COV	SD	ACC	SD	COV	SD
lymphoma	0.802	0.046	0.943	0.01	**0.891**	0.045	1.0	0.0
leukemia	0.824	0.037	1.0	0.0	**0.901**	0.021	1.0	0.0
colon	0.739	0.046	1.0	0.0	**0.818**	0.03	1.0	0.0
lung	0.918	0.013	1.0	0.0	**0.94**	0.012	1.0	0.0
prostate	0.807	0.042	1.0	0.0	**0.868**	0.015	1.0	0.0
ovarian	0.981	0.003	1.0	0.0	**0.985**	0.008	1.0	0.0
audiology	**0.524**	0.012	0.955	0.012	0.515	0.018	1.0	0.002
biodeg	0.808	0.004	1.0	0.0	**0.821**	0.008	1.0	0.0
sonar	0.755	0.022	1.0	0.0	**0.762**	0.022	1.0	0.0
cylinder	**0.682**	0.015	0.86	0.01	0.679	0.01	1.0	0.0
dermatology	0.921	0.007	1.0	0.0	**0.933**	0.007	1.0	0.0
mushroom	**1.0**	0.0	1.0	0.0	**1.0**	0.0	1.0	0.0
flags	**0.578**	0.019	1.0	0.0	**0.578**	0.024	1.0	0.0
ozone	0.948	0.004	0.825	0.003	**0.961**	0.002	1.0	0.0
parkinsons	0.88	0.019	1.0	0.0	**0.892**	0.014	1.0	0.0
SAheart	0.638	0.018	1.0	0.0	**0.646**	0.016	1.0	0.0
segmentation	**0.955**	0.002	1.0	0.0	0.949	0.003	1.0	0.0
spam	0.896	0.002	1.0	0.0	**0.899**	0.003	1.0	0.0

Tablica 9.28: Średnie ACC i COV z odchyleniami standardowymi (SD) dla zbiorów klasyfikowanych *V-drzewem* z użyciem miary *DiscPairs* za pomocą 10-krotnej CV.

decyzyjnych (P).

Ponadto w Tab. 9.31 przedstawione jest porównanie stosowanych miar w klasycznym drzewie i V-drzewie pod kątem otrzymanego ACC. Można zauważyć, że np. w przypadku zbioru "leukemia" najwyższą dokładność uzyskano dla miary *DiscPairs* i klasycznego drzewa, natomiast najniższy wynik dotyczył miary opartej na entropii. Z drugiej strony, dla *V-drzewa*, najwyższy wynik uzyskano dla miary *Entropia*, na drugim miejscu była miara *DiscPairs*, a najniższy wynik dała miara *Gini*.

Ponadto, Rys. 9.12 przedstawia porównanie obu drzew ze względu na zastosowaną miarę jakości cięć. Łatwo zauważyć, że w większości przypadków lepsze wyniki uzyskano dla klasyfikatora opartego na drzewie decyzyjnym z cięciami weryfikującymi. Ponadto, wyniki były podobne, niezależnie od wybranej miary.

V-drzewo decyzyjne utworzone dla zbioru danych $HOLTER\_II$ przedstawia Rys. 9.13. W każdym węźle drzewa podano cięcie główne oraz cięcia weryfikujące, o ile istniały takie o odpowiednio dobrej jakości. Podobnie do drzew klasycznych,

Zbiór	Klasyfikator *CTree-Entropy*				Klasyfikator *VTree-Entropy*			
	ACC	SD	COV	SD	ACC	SD	COV	SD
lymphoma	0.788	0.041	0.945	0.022	**0.836**	0.043	1.0	0.0
leukemia	0.803	0.037	1.0	0.0	**0.91**	0.023	1.0	0.0
colon	0.75	0.045	1.0	0.0	**0.756**	0.033	1.0	0.0
lung	0.925	0.014	1.0	0.0	**0.957**	0.013	1.0	0.0
prostate	0.837	0.026	1.0	0.0	**0.876**	0.024	0.999	0.002
ovarian	0.976	0.004	1.0	0.0	**0.981**	0.004	0.999	0.002
audiology	**0.625**	0.025	0.74	0.017	0.538	0.039	0.996	0.004
biodeg	0.817	0.009	1.0	0.0	**0.818**	0.009	1.0	0.0
sonar	0.74	0.03	1.0	0.0	**0.752**	0.024	1.0	0.0
cylinder	0.708	0.015	0.813	0.011	**0.73**	0.013	1.0	0.0
dermatology	0.945	0.007	1.0	0.001	**0.954**	0.008	1.0	0.0
mushroom	**1.0**	0.0	1.0	0.0	0.985	0.0	1.0	0.0
flags	0.629	0.023	1.0	0.0	**0.632**	0.019	0.999	0.002
ozone	0.953	0.003	0.843	0.004	**0.96**	0.002	1.0	0.0
parkinsons	0.865	0.016	1.0	0.0	**0.873**	0.029	1.0	0.0
SAheart	0.626	0.013	1.0	0.0	**0.647**	0.013	1.0	0.001
segmentation	**0.953**	0.002	1.0	0.0	0.945	0.002	1.0	0.0
spam	**0.921**	0.002	1.0	0.0	0.915	0.003	1.0	0.0

Tablica 9.29: Średnie ACC i COV z odchyleniami standardowymi (SD) dla zbiorów klasyfikowanych *V-drzewem* z użyciem miary *Entropia* za pomocą 10-krotnej CV.

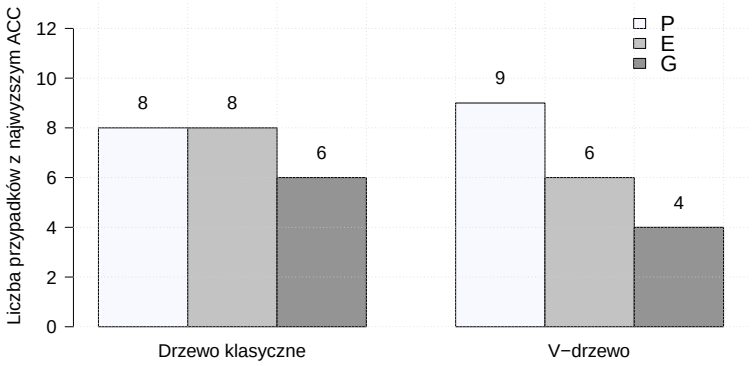
V-drzewo jako najlepiej dyskryminujące cięcie uznało to określone na atrybucie *mi* (liczba przebytych zawałów serca). Żadne inne cięcie nie uzyskało podobnie dobrej jakości, dlatego w korzeniu nie ma cięć weryfikujących.

Możliwość bezpośredniego porównywania drzew uzyskanych dla zbiorów $HOLTER\_I$ oraz $HOLTER\_II$, jest ograniczona ze względu na zastosowanie odmiennych aparatów EKG Holter dla każdego z tych zbiorów (Aspel oraz BTL). Systemy te eksportują co prawda informacje na temat takich samych pojęć, np. dotyczących zaburzeń rytmu, zmian odcinka ST czy zmienności HRV, jednak szczegółowe parametry zapisu EKG są nieznacznie odmienne. Na przykład, system Aspel w zastosowanej wersji nie eksportuje analiz odstępu PQ, dostępnych w BTL, zawiera natomiast analizę dyspersji odstępu QT, która nie jest dostępna w plikach eksportowanych przez zastosowany system BTL. W zakresie analizy widma zmienności rytmu serca HRV, eksportowane z BTL pliki nie zawierały parametrów dotyczących pasma VLF o bardzo niskiej częstotliwości (<0.0033) ani ULF o częstotliwości 0.0033 – 0.04 Hz. To może być m.in. przyczyną odmiennych

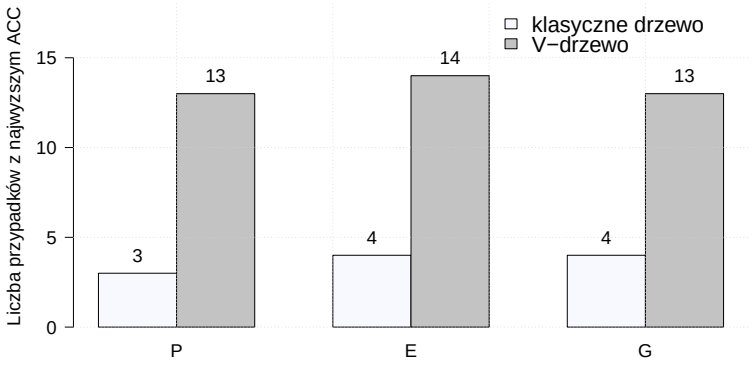
Zbiór	Klasyfikator *CTree-Gini*				Klasyfikator *VTree-Gini*			
	ACC	SD	COV	SD	ACC	SD	COV	SD
lymphoma	0.795	0.042	0.943	0.021	**0.845**	0.047	1.0	0.0
leukemia	0.819	0.035	1.0	0.0	**0.9**	0.04	1.0	0.0
colon	**0.766**	0.03	1.0	0.0	0.765	0.045	1.0	0.0
lung	0.925	0.014	1.0	0.0	**0.956**	0.02	1.0	0.0
prostate	0.84	0.033	1.0	0.0	**0.847**	0.014	1.0	0.0
ovarian	0.976	0.004	1.0	0.0	**0.98**	0.006	1.0	0.0
audiology	**0.66**	0.02	0.827	0.026	0.618	0.021	1.0	0.0
biodeg	0.809	0.008	1.0	0.0	**0.813**	0.011	1.0	0.0
sonar	0.695	0.02	1.0	0.0	**0.722**	0.024	1.0	0.0
cylinder	0.703	0.014	0.811	0.014	**0.74**	0.015	1.0	0.0
dermatology	0.939	0.005	0.998	0.001	**0.952**	0.006	1.0	0.0
mushroom	**1.0**	0.0	0.787	0.0	**1.0**	0.0	1.0	0.0
flags	0.605	0.017	1.0	0.0	**0.609**	0.019	1.0	0.0
ozone	0.947	0.004	0.822	0.002	**0.96**	0.003	1.0	0.0
parkinsons	0.86	0.025	1.0	0.0	**0.882**	0.025	1.0	0.0
SAheart	0.613	0.009	1.0	0.0	**0.652**	0.015	1.0	0.001
segmentation	**0.955**	0.003	1.0	0.0	0.942	0.003	1.0	0.0
spam	**0.913**	0.002	1.0	0.0	0.893	0.003	1.0	0.0

Tablica 9.30: Średnie ACC i COV z odchyleniami standardowymi (SD) dla zbiorów klasyfikowanych *V-drzewem* z użyciem miary *Gini* za pomocą 10-krotnej CV.

drzew decyzyjnych generowanych dla zbioru $HOLTER\_I$ oraz $HOLTER\_II$, przy czym zachowane jest podobieństwo drzew dla poszczególnych zbiorów między różnymi metodami.



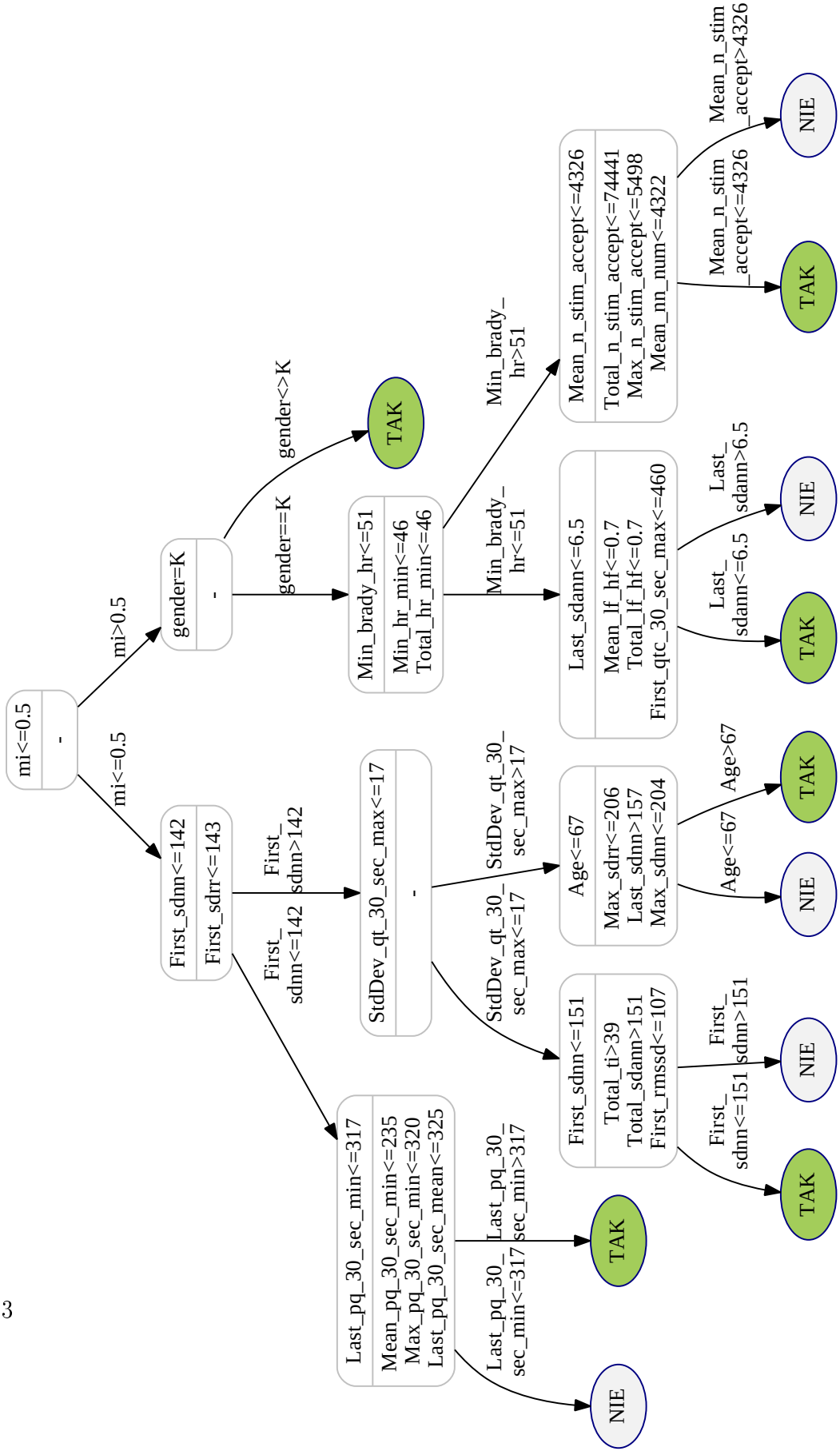
Rysunek 9.11: Porównanie miar jakości cięć.



Rysunek 9.12: Porównanie drzew.

Zbiór	*Drzewo klasyczne*			*V-drzewo*		
	1-wszy	2-gi	3-ci	1-wszy	2-gi	3-ci
lymphoma	P	G	E	P	G	E
leukemia	P	G	E	E	P	G
colon	G	E	P	P	G	E
lung	E, G	P	-	E	G	P
prostate	G	E	P	E	P	G
ovarian	P	E, G	-	P	E	G
audiology	G	E	P	G	E	P
biodeg	E	G	P	P	E	G
sonar	P	E	G	P	E	G
cylinder	E	G	P	G	E	P
dermatology	E	G	P	E	G	P
mushroom	P, E, G	-	-	P, G	E	-
flags	E	G	P	E	G	P
ozone	E	P	G	P	E,G	-
parkinsons	P	E	G	P	G	E
SAheart	P	E	G	G	E	P
segmentation	P, G	E	-	P	E	G
spam	E	G	P	E	P	G

Tablica 9.31: Porównanie miar zastosowanych w klasycznym drzewie i V-drzewie względem uzyskanych wartości ACC (P oznacza *DiscPairs*, G oznacza *Gini* oraz E oznacza *Entropię*).



Rysunek 9.13: V-drzewo dla zbioru danych *HOLTER_II*.

9.4.2 Statystyczna weryfikacja hipotez dotyczących V-drzewa

Celem sprawdzania istotności różnic między wynikami klasyfikacji uzyskanymi za pomocą klasycznego drzewa decyzyjnego (*CTree-Disc*) oraz drzewa z cięciami weryfikującymi (*VTree-Disc*) przeprowadzono test Wilcoxona dla par obserwacji zależnych. Test ten stanowi nieparametryczną alternatywę dla testu t-Studenta, jednak w przeciwieństwie do testu t-Studenta, nie posiada założeń dotyczących rozkładu próby. Test Wilcoxona dla par obserwacji stosowany jest do porównania wartości danej cechy w parach, gdzie jedna z wartości w parze pochodzi z populacji X, a druga z populacji Y. Na podstawie n par obserwacji $(x_1, y_1), \ldots, (x_n, y_n)$ można ocenić, czy populacje X i Y mają takie same rozkłady. W badaniach zastosowana została wersja testu Wilcoxona zaimplementowana w programie *STATISTICA* [146]. Za pomocą tego testu zweryfikowano hipotezę zerową postaci: jakość klasyfikacji jest jednakowa dla obu metod (H_0) na poziomie istotności równym 0.05.

Tab. 9.32 zawiera wszystkie hipotezy statystyczne, które były przedmiotem weryfikacji, jak również wyniki wykonywanych testów. Wyniki są przedstawione w postaci minimalnej wartości granicznego poziomu istotności, tj. wartości p, dla której hipoteza zerowa powinna zostać odrzucona. Łatwo zauważyć, że niemal wszystkie są znacząco mniejsze od 0.05. Dodatkowo wszystkie wyniki o wartości parametru p mniejszej od 0.05 (wymaganej do odrzucenia hipotezy H_0 w tym teście) są przedstawione w postaci wykresów pudełkowych (ang. *box-and-whisker plots, boxplots*) ze wskazaniem mediany, 25-go i 75-go percentyla (I i III kwartyla) oraz wartości minimalnych i maksymalnych w postaci tzw. *wąsów*. Wykresy przedstawia Tab. 9.33. Porównują one jakość klasyfikacji przeprowadzoną przez *V-drzewa* (*VTree-Disc*) oraz drzewa klasyczne (*CTree-Disc*). Wyniki testu wskazują, że *VTree-Disc* daje statystycznie istotnie lepsze wyniki niż *CTree-Disc*.

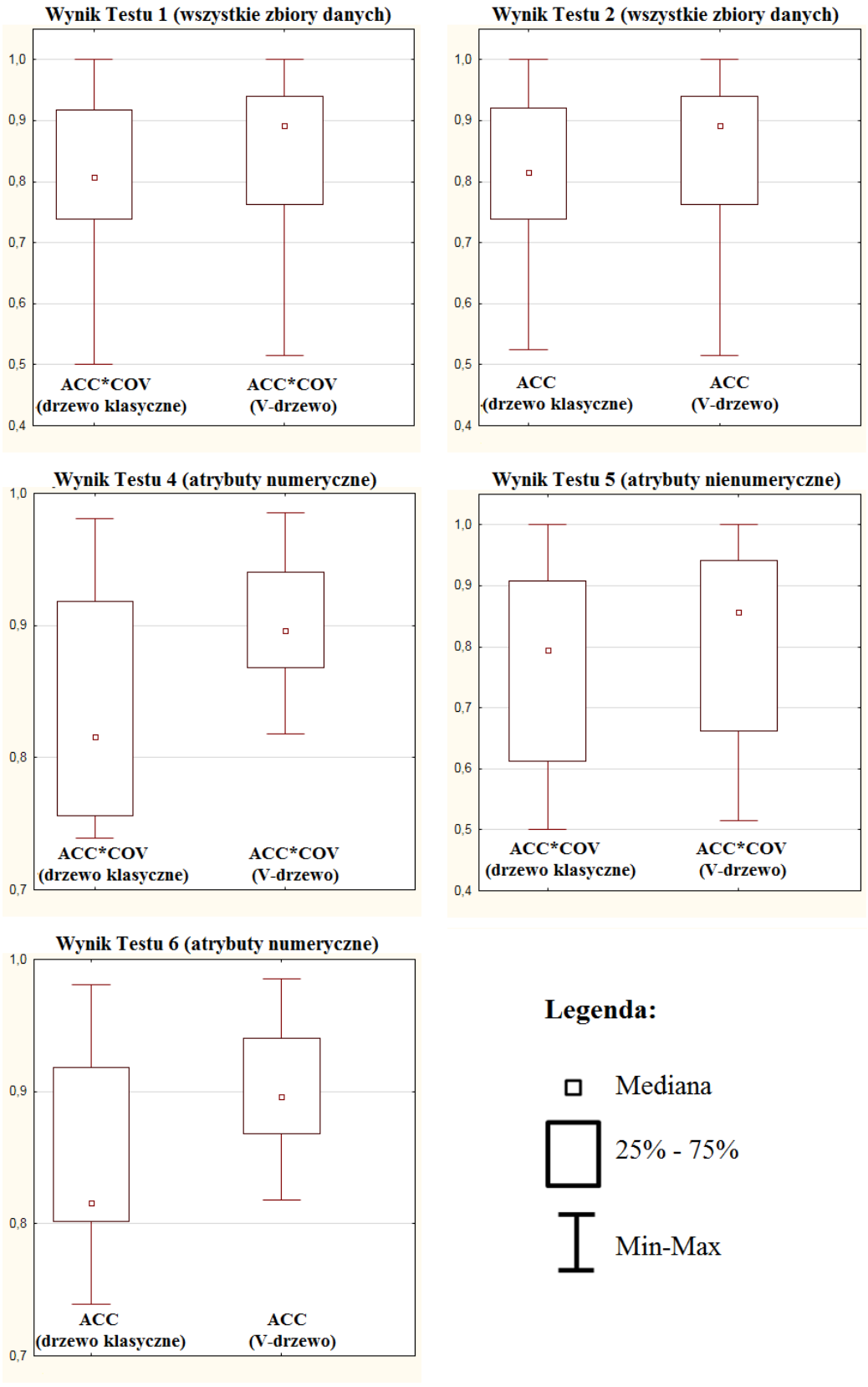
Przeprowadzone testy statystyczne potwierdzają użyteczność *V-klasyfikatora* w zagadnieniu klasyfikacji. Ten nowy typ klasyfikatora działa statystycznie lepiej niż wersja klasyczna, jeżeli porównujemy dwie miary jakości klasyfikacji, to jest ACC i $ACC \cdot COV$). W przypadku zastosowania współczynnika pokrycia, algorytm *V-drzewa* nigdy nie klasyfikował gorzej od klasyfikatora klasycznego, jednak obserwacja nie jest statystycznie istotna ponieważ wartości współczynnika pokrycia otrzymane dla *V-drzewa* oraz drzewa klasycznego są różne tylko dla 4 zbiorów danych (ze wszystkich 18 zbiorów). Należy także zaobserwować, że *V-klasyfikator* miał przewagę w porównaniu do klasycznego w odniesieniu do współczynnika dokładności klasyfikacji oraz zbiorów danych z dużą liczbą atrybutów (Tab. 9.7). Takie spostrzeżenie nie jest prawdziwe dla zbiorów danych z liczbą atrybutów nieprzekraczającą kilkuset (chociaż takie dane przeważają w przeprowadzonych eksperymentach nad danymi z wieloma atrybutami).

Wyniki przeprowadzone na 18 zbiorach danych potwierdzają, że wykorzystanie

Numer testu	Hipoteza zerowa H_0	Progowe wartości wsp. istotności wystarczające do odrzucenia H_0
1	Jakość klasyfikacji wyrażona jako iloczyn **Acc·Cov** dla V-drzewa i drzewa klasycznego jest równa	0.0007
2	Jakość klasyfikacji wyrażona jako wartość **Acc** dla V-drzewa i drzewa klasycznego jest równa	0.0037
3	Jakość klasyfikacji wyrażona jako wartość **Cov** dla V-drzewa i drzewa klasycznego jest równa	0.0678
4	Jakość klasyfikacji wyrażona jako iloczyn **Acc·Cov** dla V-drzewa i drzewa klasycznego oraz dla danych z wieloma atrybutami jest równa	0.0277
5	Jakość klasyfikacji wyrażona jako iloczyn **Acc·Cov** dla V-drzewa i drzewa klasycznego oraz dla danych z niezbyt licznym zbiorem atrybutów jest równa	0.0093
6	Jakość klasyfikacji wyrażona jako wartość **Acc** dla V-drzewa i drzewa klasycznego oraz dla danych z wieloma atrybutami jest równa	0.0277
7	Jakość klasyfikacji wyrażona jako wartość **Acc** dla V-drzewa i drzewa klasycznego oraz dla danych z niezbyt licznym zbiorem atrybutów jest równa	0.0744

Tablica 9.32: Wyniki testu Wilcoxona dla par obserwacji.

dodatkowej wiedzy dziedzinowej zawartej w atrybutach redundantnych poprawia jakość klasyfikacji. Wyniki eksperymentów wskazują, że w większości przypadków zbiorów danych dokładność przedstawionego klasyfikatora była lepsza w porównaniu do klasycznego. Wzrost dokładności wynosił od nieznaczącej wartości 0.3% do aż 9%. Zaobserwować można także lepsze wyniki (wzrost ACC o 8.9%, 7.7% czy 7.9%) dla danych mikromacierzowych (6 z 18 zbiorów danych), które charakteryzują się bardzo dużą liczbą atrybutów oraz małą liczbą obiektów.



Tablica 9.33: Wykresy pudełkowe testów z poziomem istotności poniżej 0.05.

9.5 Eksperymenty z odległością ontologiczną

W eksperymentach z odległością ontologiczną dla danych ze zbioru $HOLTER\_I$ wykorzystano opisaną w Rozdziale 7.1 ontologię choroby niedokrwiennej serca oraz dane pacjentów z tą chorobą o różnym stopniu nasilenia. Do badań zbioru $HOLTER\_II$ ontologię zmodyfikowano celem dostosowania pojęć do atrybutów dostępnych w zbiorze. Wynika to z zastosowania odmiennych aparatów EKG do zapisu holterowskiego, które generują nieznacznie odmienne parametry. Ontologię wykorzystaną do eksperymentów z danymi ze zbioru $HOLTER\_II$ przedstawia Rys. 9.14.

Celem oceny efektywności odległości ontologicznej wyznaczonej według zaproponowanego w Rozdz. 7.2 algorytmu przeprowadzono 4 rodzaje testów: E1, E2, E3, E4, scharakteryzowane w Tab. 9.34 oraz 9.35 odpowiednio dla zbioru $HOLTER\_I$ i $HOLTER\_II$. Do testów wykorzystano metodę k-NN zaimplementowaną w systemie WEKA [156] z własną modyfikacją dla odległości ontologicznej oznaczoną dalej jako *kNN-OntoDist*. Do reprezentacji modelu ontologii wykorzystano bibliotekę Java o nazwie SOFA (*Simple Ontology Framework API* [142]). Doświadczenia przeprowadzono z parametrem k równym 3 w przypadku zbioru $HOLTER\_I$ oraz 5 dla $HOLTER\_II$. Schematy poszczególnych testów różniły się zastosowaną odległością (Euklidesa lub ontologiczna wyznaczona na podstawie ontologii z Rys. 7.1 lub 9.14, zawierającej 31 pojęć) oraz sposobem wyznaczania wag pojęć w ontologii (wskazane przez eksperta lub wygenerowane losowo metodą Monte Carlo). Do oceny jakości klasyfikacji w przypadku zbioru $HOLTER\_I$ zastosowano technikę LOO oraz 10-CV dla $HOLTER\_II$, z wyjątkiem ostatniego eksperymentu (E4), gdzie była to zagnieżdżona walidacja krzyżowa (ang. *nested-CV*). W technice zagnieżdżonej, walidację zewnętrzną przeprowadzono metodą LOO ($HOLTER\_I$) oraz 10-CV ($HOLTER\_II$). W każdym zbiorze treningowym generowano 100 modeli ontologii z losowymi wagami i wybierano najlepszy model (o największym ACC) techniką 10-CV do testowania zewnętrznego. Schemat przeprowadzonej zagnieżdżonej walidacji krzyżowej przedstawia Rys. 9.15.

Ostateczny wynik jest średnią wszystkich testów. Charakterystykę eksperymentów oraz ich wyniki przedstawia Tab. 9.34 dla zbioru $HOLTER\_I$ oraz Tab. 9.35 dla zbioru $HOLTER\_II$.

Dla obydwu zbiorów widoczna jest znaczna rozbieżność między dokładnością metody k_NN z odległością Euklidesa a odległością ontologiczną, na korzyść tej drugiej. W zbiorze $HOLTER\_I$ interesująca jest niewielka różnica ACC pomiędzy odległością ontologiczną z wagami wskazanymi przez eksperta a tą odległością z wagami generowanymi losowo. Sugeruje to, że znacznie większe znaczenie ma dobór pojęć do ontologii zgodnie w wiedzą dziedzinową, niż wagi poszczególnych pojęć ontologii. Jednak odpowiedni dobór wag może jeszcze, poza odpowiednią selekcją pojęć, poprawić dokładność klasyfikacji, na co wskazuje wynik ekspery-

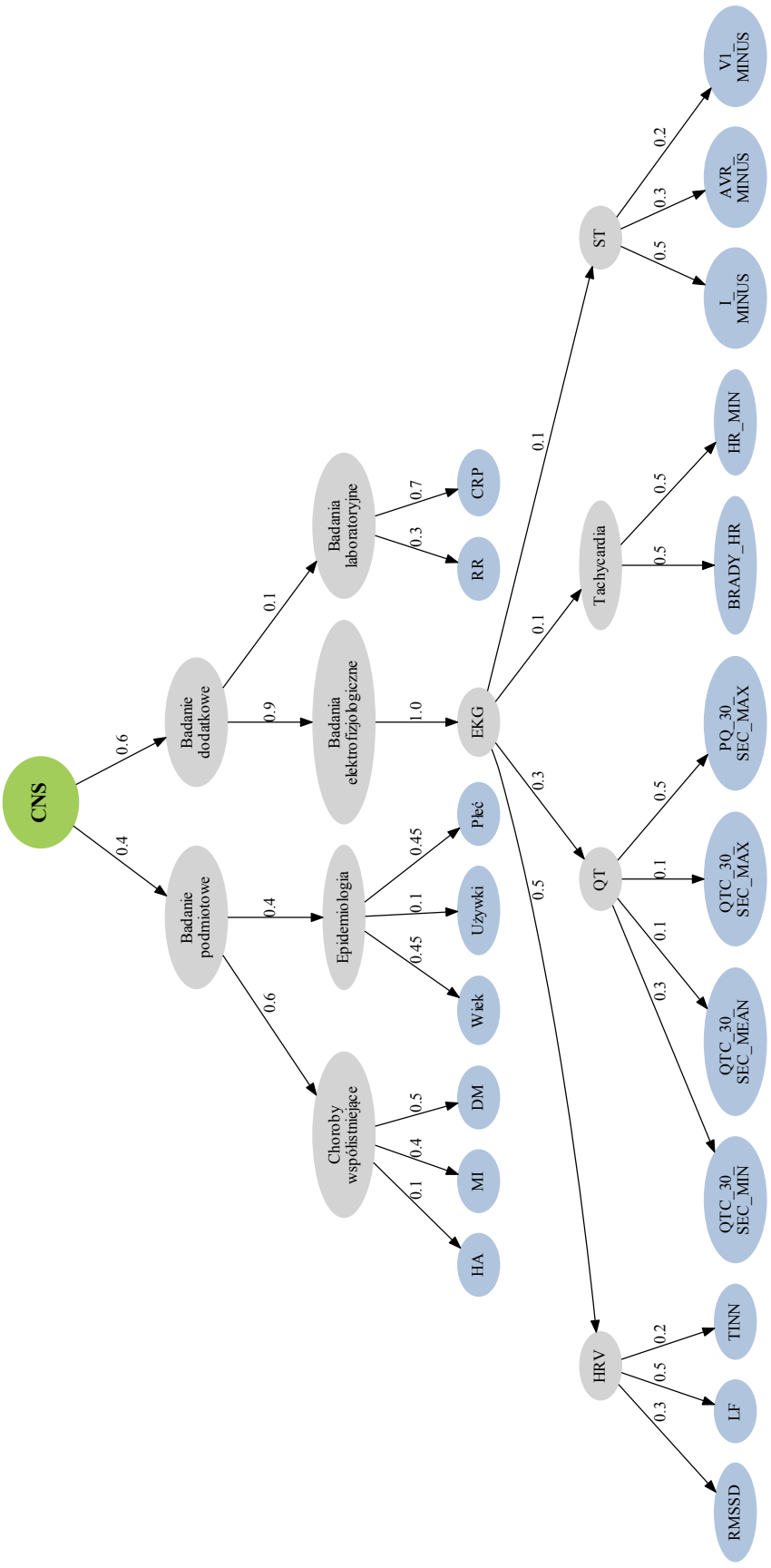
Cecha	**E1**	**E2**	**E3**	**E4**
Odległość	Euklidesa	Ontologiczna	Ontologiczna	Ontologiczna
Wagi pojęć	-	eksperta	Monte Carlo	Monte Carlo
Ocena jakości	LOO	LOO	LOO	zagnieżdżona CV
Dokładność	82.35%	94.12%	93.52% (średnia z 12 powtórzeń, SD=0.0475)	98.53%

Tablica 9.34: Wyniki eksperymentów z wykorzystaniem metody IV do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_I$.

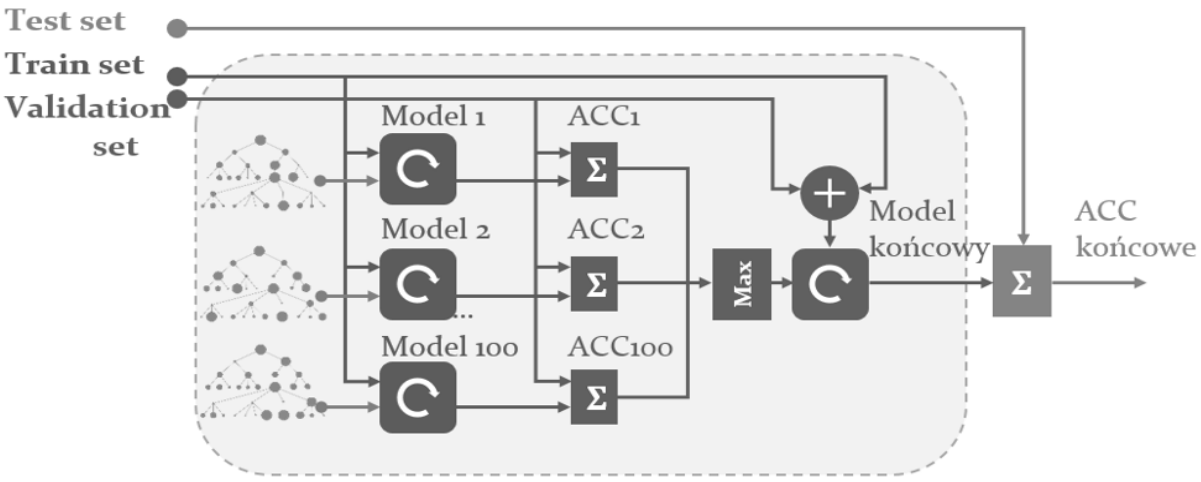
Cecha	**E1**	**E2**	**E3**	**E4**
Odległość	Euklidesa	Ontologiczna	Ontologiczna	Ontologiczna
Wagi pojęć	-	eksperta	Monte Carlo	Monte Carlo
Ocena jakości	10-CV	10-CV	10-CV	zagnieżdżona CV
Dokładność	68.50%	92.50%	84.5% (średnia z 12 powtórzeń, SD=0.08)	93%

Tablica 9.35: Wyniki eksperymentów z wykorzystaniem metody IV do predykcji stenoz wieńcowych w CNS dla zbioru $HOLTER\_II$.

mentu E4, w którym wagi są wielokrotnie losowane, a do modelu wybierane są tylko najlepsze z nich. Taki sposób doboru wag, ważący pojęcia lepiej niż ekspert, może wynikać z trudności bardzo dokładnego numerycznego oszacowania wag przez człowieka jednocześnie dla wielu (tutaj 31) pojęć w ontologii.



Rysunek 9.14: Ontologia CNS z wagami wskazanymi przez eksperta dla zbioru *HOLTER_II*.



Rysunek 9.15: Schemat zagnieżdżonej walidacji krzyżowej.

9.6 Eksperymenty z metodą mierzenia wpływu czynnika na percepcję

9.6.1 Drzewo wpływu i reguły krzyżowe

Celem prezentacji *drzewa wpływu* (*I-drzewa*) dla danych rzeczywistych przeprowadzono eksperymenty z użyciem danych $HOLTER\_I$ dla wszystkich 70 pacjentów. Czynnikiem zakłócającym percepcję była ponadstandardowa terapia lekiem Z (zileuton) o działaniu przeciwzapalnym, którą zastosowano u połowy pacjentów (53%).

I-drzewo utworzone dla powyższych danych prezentuje Rys. 8.1. W każdym węźle przedstawiono liczbę pacjentów otrzymujących placebo bez istotnych stenoz ($P0$), z jednym istotnym zwężeniem ($P1$), dwoma ($P2$) i trzema zwężeniami ($P3$), jak i tych leczonych zileutonem bez istotnych zwężeń ($Z0$), z jednym ($Z1$), dwoma ($Z2$) i trzema istotnymi stenozami ($Z3$). Przedstawiono także wartość oczekiwaną liczby stenoz (S), osobno w grupie otrzymującej placebo ($E(S|P)$), jak i grupie otrzymującej zileuton ($E(S|Z)$). Dla każdego węzła wyliczono różnicę między tymi wartościami oczekiwanymi (wartość δ) oraz wartość oczekiwaną różnicy w liczbie stenoz między grupą leczoną i nieleczoną ($E(X)$).

Wszystkie obiekty należące od jednego węzła charakteryzują się takim samym wzorcem EKG, określonym przez ścieżkę od korzenia drzewa do tego węzła. Uzyskane drzewo posiada sześć liści, dostarczających po jednej *regule krzyżowej* zdefiniowanej w podrozdziale 8.2.1. Na podstawie uzyskanych reguł można stwierdzić, że nie wszyscy pacjenci reagowali identycznie na dodatkową terapię. U części pacjentów działanie leku Z było korzystne, u części obojętne, a w pewnej grupie przynosiło niekorzystne efekty w postaci zmienionego EKG.

Przykładowa *reguła krzyżowa* liścia drzewa z Rys. 8.1 opisująca korzystne działanie terapii lekiem Z ma postać:

$$\left.\begin{array}{c} AVG\_ST\_DOWN3 < -0.06 \\ \wedge FIRST\_VLF \geq 373 \\ \wedge AVG\_QT2\_AVG \geq 464.2 \end{array}\right| \Rightarrow \left|\begin{array}{l} E(S|terapia = P) = 0.29 \\ E(S|terapia = Z) = 2.33 \end{array}\right. \tag{9.1}$$

Poprzednik reguły zawiera parametry EKG określone przez ścieżkę od korzenia drzewa do tego liścia. Reguła ta mówi, że liczba istotnie zwężonych tętnic wieńcowych w grupie pacjentów nieleczonych badaną farmakoterapią wynosiła przeciętnie 0.29, a w grupie leczonej 2.33, natomiast wszyscy pacjenci (z grupy leczonej i nieleczonej) mają taki sam wzorzec EKG, tj.: średnie obniżenie odcinka ST w trzecim odprowadzeniu w ciągu doby (AVG_ST_DOWN3) nie przekraczało poziomu -0.06 mV, pierwsza w dobowym oknie wartość mocy pasma HRV (ang. *heart rate*

variability) bardzo niskiej częstotliwości - $FIRST\_VLF$ (ang. *very low frequency*) było większe niż 373 ms^2, a średni w ciągu doby ze średnich godzinowych czas trwania odstępu QT w odprowadzeniu 2 (AVG_QT2_AVG) przekraczał poziom 464 mV. Zapis EKG jest wspólny dla wszystkich obiektów należących do tego węzła: 7 pacjentów bez dodatkowej terapii i 3 otrzymujących lek Z, jednak grupa pacjentów otrzymujących zileuton ma większą liczbę zwężonych tętnic. Sposób wyliczania wartości oczekiwanych decyzji w obu powyższych grupach przedstawiają wzory 9.2 i 9.3.

$$\begin{aligned} E(S \mid terapia = P) = 0 \cdot \frac{P(S=0, terapia=P)}{P(terapia=P)} + 1 \cdot \frac{P(S=1, terapia=P)}{P(terapia=P)} + \\ + 2 \cdot \frac{P(S=2, terapia=P)}{P(terapia=P)} + 3 \cdot \frac{P(S=3, terapia=P)}{P(terapia=P)} = \\ = 0 \cdot \frac{0.6}{0.7} + 1 \cdot \frac{0}{0.7} + 2 \cdot \frac{0.1}{0.7} + 3 \cdot \frac{0}{0.7} = \\ = 0 \cdot 0.86 + 1 \cdot 0 + 2 \cdot 0.14 + 3 \cdot 0 = 0.29 \end{aligned} \tag{9.2}$$

$$\begin{aligned} E(S \mid terapia = Z) = 0 \cdot \frac{P(S=0, terapia=Z)}{P(terapia=Z)} + 1 \cdot \frac{P(S=1, terapia=Z)}{P(terapia=Z)} + \\ + 2 \cdot \frac{P(S=2, terapia=Z)}{P(terapia=Z)} + 3 \cdot \frac{P(S=3, terapia=Z)}{P(terapia=Z)} = \\ = 0 \cdot \frac{0}{0.3} + 1 \cdot \frac{0.1}{0.3} + 2 \cdot \frac{0}{0.3} + 3 \cdot \frac{0.2}{0.3} = \\ = 0 \cdot 0 + 1 \cdot 0.33 + 2 \cdot 0 + 3 \cdot 0.67 = 2.33 \end{aligned} \tag{9.3}$$

W sensie zapisu EKG pacjenci tego węzła są nierozróżnialni. Tak więc, dodatkowa terapia zmienia zapis EKG, na taki jaki występuje u pacjentów nieleczonych bez istotnych zwężeń naczyń. Rodzaj tej modyfikacji określa wartość δ (wzór 8.13), która dla tego liścia wynosi 2.05 i jest większa od założonego progu wynoszącego 1.75. Zatem dla obiektów pasujących do wzorca z powyższej reguły krzyżowej oczekujemy korzyści ze stosowania czynnika zakłócającego percepcję, czyli dodatkowej terapii lekiem Z. Taka reguła może stanowić wskazówkę do kontynuacji terapii tym lekiem.

Natomiast negatywny wpływ leczenia dodatkową farmakoterapią dotyczy obiektów należących do liścia drzewa z Rys. 8.1 opisanego regułą krzyżową po-

staci:

$$\begin{array}{c} AVG\_ST\_DOWN3 < -0.06 \\ \wedge \\ FIRST\_VLF < 373 \end{array} \Rightarrow \left| \begin{array}{l} E(S|terapia = P) = 2.17 \\ E(S|terapia = Z) = 0.2 \end{array} \right. \tag{9.4}$$

Obiekty te charakteryzują się wspólnym wzorcem, tj. średnim obniżeniem odcinka ST w trzecim odprowadzeniu w ciągu doby (AVG_ST_DOWN3) nieprzekraczającym poziomu -0.06 mV oraz pierwszej w dobowym oknie wartości mocy pasma HRV bardzo niskiej częstotliwości ($FIRST\_VLF$) mniejszej od 373 ms^2. Pomimo dodatkowego leczenia, zapis EKG jest taki sam, jak u nieleczonych pacjentów z istotnie zwężonymi naczyniami.

Przykładem reguły opisującej obiekty, u których nie obserwuje się wpływu czynnika zakłócającego jest Reguła 9.5:

$$\begin{array}{l} AVG\_ST\_DOWN3 < -0.06 \\ \quad \wedge FIRST\_VLF \geq 373 \\ \wedge AVG\_QT2\_AVG < 464.2 \\ \wedge AVG\_QT1\_STD \geq 26.14 \end{array} \Rightarrow \left| \begin{array}{l} E(S|therapy = P) = 0.25 \\ E(S|therapy = Z) = 1.5 \end{array} \right. \tag{9.5}$$

gdzie AVG_QT2_AVG to średni dobowy ze średnich godzinowych czas trwania odstępu QT w drugim odprowadzeniu, a AVG_QT1_STD oznacza średnie w ciągu doby odchylenie standardowe z godzinowych pomiarów odstępu QT. Dla tej reguły, różnica w liczbie stenoz pomiędzy grupą leczoną i nieleczoną lekiem Z jest zbyt mała ($(E(X_{\mathbf{A}}) = 1.33) < 1.75$), aby stwierdzić wpływ czynnika zakłócającego. Powodem takiego stanu rzeczy może być zbyt mała liczba obiektów uniemożliwiająca dalsze podziały lub rzeczywisty brak działania dodatkowego leczenia. W tej ostatniej sytuacji można byłoby mówić o neutralnym liściu. Z praktycznego punktu widzenia dla obiektów z neutralnych liści należałoby rozważyć zaprzestanie podawania dodatkowego leku ze względu na brak korzystnego efektu w zapisie EKG.

9.6.2 Statystyczna weryfikacja hipotez dotyczących I-drzewa

Celem sprawdzenia statystycznej istotności różnic w rozkładach wartości decyzji, tj. liczby stenoz $(0, 1, 2, 3)$ w obu grupach (placebo i zileuton) wykonano test χ^2 dla zmiennych jakościowych dla każdego liścia drzewa wpływu z Rys. 8.1. Test weryfikuje hipotezę zerową (H_0), która mówi, że frakcje w obu grupach są takie same na poziomie istotności $\alpha = 0.05$. Testem tym porównano również osobno każdą z wartości $0, 1, 2, 3$, jak i wszystkie razem w obu grupach, celem sprawdzenia czy

ich liczebności są różne w obu grupach. Wyniki testów statystycznych przedstawia Tab. 9.36. Numeracja liści z Rys. 8.1 odbywa się poziomami z góry na dół i od strony lewej do prawej (wg porządku *level-order* dotyczącym tylko liści drzewa).

Stenozy w liściach		Grupa		χ^2	p	p dla porównań ilości 0,1,2,3 osobno
		Placebo	Zileuton			
Liść 1	0	0%	100%	9	**0.001**	0.0015
	2	20%	0%			0.17
	3	80%	0%			0.008
	All	55.56%	44.44%			**0.38**
Liść 2	0	16.67%	80%	10	**0.04**	0.017
	1	0%	20%			0.12
	2	33.33%	0%			0.09
	3	50%	0%			0.03
	All	54.55%	45.45%			**0.4**
Liść 3	0	85.71%	0%	10	**0.02**	0.01
	1	0%	33.33%			0.06
	2	14.29%	0%			0.25
	3	0%	66.67%			0.006
	All	70%	30%			**0.12**
Liść 4	0	75%	16.67%	4	0.3	0.03
	1	25%	33.33%			0.39
	2	0%	33.33%			0.1
	3	0%	16.67%			0.2
	All	40%	60%			**0.27**
Liść 5	0	25%	58.33%	3.5	0.17	0.11
	1	25%	33.33%			0.38
	2	50%	8.33%			0.03
	All	25%	75%			0.04
Liść 6	0	57.14%	33.33%	0.7	0.4	0.19
	1	42.86%	66.67%			0.19
	All	53.85%	46.15%			**0.19**

Tablica 9.36: Wyniki testów statystycznych dla każdego liścia *drzewa wpływu* z Rysunku 8.1.

We wszystkich liściach, z wyjątkiem piątego, liczba osób z placebo nie różni się od liczby osób otrzymujących zileuton. Wskazują na to wartości p (wyróżnione pogrubioną czcionką w ostatniej kolumnie Tab. 9.36) przekraczające założoną wartość istotności statystycznej (0.05). W związku z tym, uzasadnione jest przeprowadze-

nie dalszego testu na równość rozkładów.

Statystycznie istotne różnice w rozkładach wartości decyzji $(0, 1, 2, 3)$ pomiędzy grupą z placebo i grupą z zileutonem występują w następujących liściach:

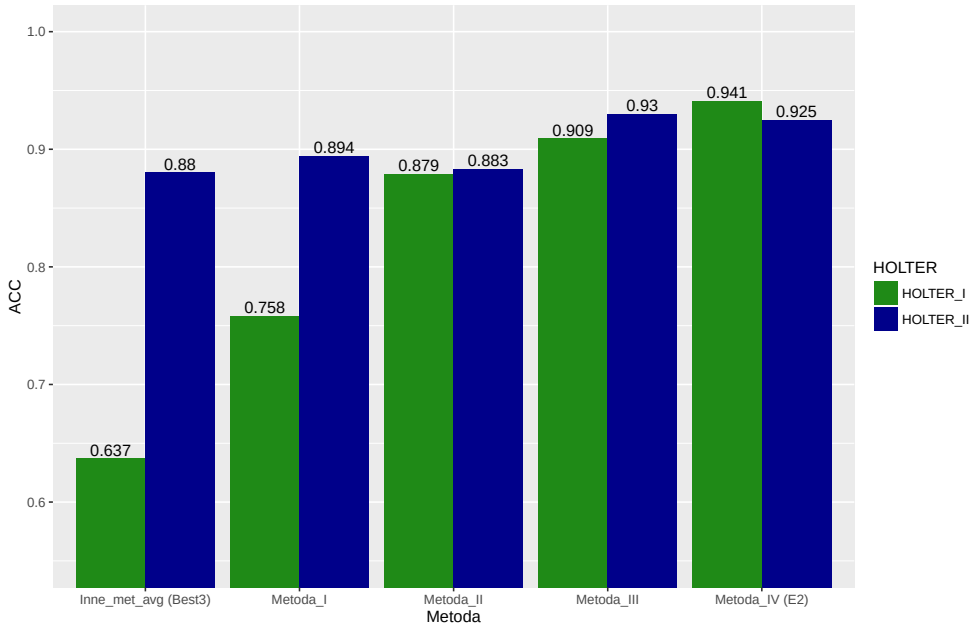
- Nr 1 z $E(X) = 2.8,\ p = 0.001$: liczba zmienionych naczyń w grupie zileutonu jest na poziomie 0, natomiast w grupie placebo między 2 a 3;
- Nr 2 z $E(X) = 2.03,\ p = 0.04$: liczba zmienionych naczyń w grupie zileutonu wynosi około 0, w grupie placebo natomiast około 3;
- Nr 3 z $E(X) = 2.14,\ p = 0.02$: liczba zmienionych naczyń w grupie zileutonu jest na poziomie 3, w grupie placebo na poziomie 0.

Takie wyniki świadczą o tym, że w tych trzech liściach, dodatkowa terapia lekiem Z, wywołuje statystycznie istotne zmiany (wartości p są poniżej poziomu istotności 0.05 - kolumna p w Tab. 9.36). Wiadomo zatem, że zmiany są znamienne, natomiast kierunek tych zmian (korzystne lub niekorzystne) można określić za pomocą wartości δ w sposób opisany w Rozdziale 8.2.3 (wzór 8.14) .

Podsumowując, w analizowanej grupie, znaczący wpływ terapii zileutonem obserwuje się u 33.3% pacjentów leczonych tym lekiem. Ta informacja może być wykorzystana przy podejmowaniu decyzji dotyczącej dalszego leczenia tych pacjentów, wskazując czy warto kontynuować u nich terapię zileutonem, czy też leczenie powinno zostać przerwane, aby nie generować niepotrzebnych kosztów oraz nie narażać pacjentów na ewentualne efekty uboczne towarzyszące dodatkowemu leczeniu.

9.7 Zestawienie głównych wyników badań

Celem sformułowania jednoznacznych wniosków końcowych sporządzono zestawienie głównych wyników badań umożliwiających porównanie tych wyników, ich analizę i uogólnienia. Rysunek 9.16 prezentuje sumaryczne zestawienie wyników eksperymentalnych.



Rysunek 9.16: Zestawienie wyników w postaci dokładności klasyfikacji (ACC) zaproponowanych metod w predykcji stenoz wieńcowych dla zbioru $HOLTER\_I$ oraz $HOLTER\_II$.

W przypadku obydwu zbiorów: $HOLTER\_I$ oraz $HOLTER\_II$, proponowane metody budowy modeli klasyfikatorów z wykorzystaniem wiedzy dziedzinowej WD osiągają lepszą dokładność klasyfikacji niż metody klasyczne, tj. C4.5, NB, SVM, k-NN, RandomForest, ANN i ModLEM, bez dodatkowej WD. Średnia dokładność trzech najlepszych spośród metod klasycznych wynosi 63.7% dla zbioru $HOLTER\_I$ oraz 88% dla $HOLTER\_II$. Metody proponowane w rozprawie do budowy drzew decyzyjnych z dodatkową WD osiągają średnią dokładność wynoszącą 84.9% oraz 90.2%, odpowiednio. Dla danych ze zbioru $HOLTER\_I$, dokładność klasyfikacji wszystkich proponowanych metod klasyfikacji (metody I, II, III i IV) wynosiła średnio 87.2%, natomiast dla zbioru $HOLTER\_II$ 90.8%.

Wyniki eksperymentów wskazują, że proponowane metody przewidywania obecności zwężeń tętnic wieńcowych wykorzystujące dodatkową wiedzę dziedzinową dają dobre wyniki, porównywalne lub lepsze od wyników innych metod. Na-

leży podkreślić, że metody II (*CTree-DiscW*) i III (*VTree-Disc*) uzyskały czułość na poziomie 94.4%, co jest bardzo pożądaną właściwością w zastosowaniach medycznych, gdyż odsetek pacjentów pozostawionych bez leczenia a wymagających interwencji powinien być minimalizowany (liczba przypadków fałszywie ujemnych FN powinna być jak najmniejsza).

W kolejnych zestawieniach przedstawiono średnie wartości dokładności klasyfikacji dla zbiorów medycznych oraz osobno dla 18 zbiorów z ogólnie dostępnych repozytoriów z wyszczególnieniem wyników każdej z proponowanych metod. Jako punkt odniesienia zastosowano średnią dokładność klasyfikacji innych testowanych metod (NB, DT, DR, ANN, k-NN, SVM i RF) niewykorzystujących dodatkowej wiedzy dziedzinowej (WD). Tab. 9.37 przedstawia średnie ACC dla obydwu zbiorów medycznych $HOLTER\_I$ oraz $HOLTER\_II$.

Metoda	**Opis metody**	**ACC**
Inne metody bez dodatkowej WD	NB, DT, DR, ANN, k-NN, SVM, RF	0.685
Metoda I (*CTree-Disc)*	DT - wzorce czasowe	0.826
Metoda II (*CTree-DiscW)*	DT - modyfikacja jakości podziałów	0.881
Metoda III (*VTree-Disc)*	DT - cięcia weryfikujące (V-drzewo)	0.920
Metoda IV (*kNN-OntoDist)*	k-NN - odległość ontologiczna	0.927
I_II_III	DT z WD	0.876
I_II_III_IV	Metody z WD	0.888

Tablica 9.37: Średnia dla danych medycznych ($HOLTER\_I$ i $HOLTER\_II$) dokładność predykcji stenoz wieńcowych w CNS.

Podsumowanie wyników dla 18 dodatkowych zbiorów danych i 3 badanych miar jakości cięć zawiera Tab. 9.38.

Dodatkowy problem medyczny: "Zdrowe serce"

Ze względu na zainteresowanie ze strony klinicystów możliwością rozpoznawania stanu tzw. "zdrowego serca" wykonano dodatkowo serię eksperymentów z wykorzystaniem proponowanych metod do identyfikacji tego stanu. Doświadczenia przeprowadzono z użyciem metod odpowiednich dla sformułowanego problemu oraz tylko dla danych ze zbioru $HOLTER\_II$, ze względu na dostęp w nich do szczegółowych informacji o procentowych zwężeniach poszczególnych naczyń. Przeprowadzone eksperymenty mają na celu potwierdzenie skuteczności metod zaproponowanych w rozprawie. "Zdrowe serce" oznacza taki stopień nasilenia choroby niedokrwiennej serca, w którym standardowo wystarczające jest leczenie zachowawcze. Ma więc

Metoda	**Miara jakości cięć**	**ACC**
Metoda I (*CTree-Disc)*	DiscPairs	0.814
Metoda I (*CTree-Entropy)*	Entropia	0.826
Metoda I (*CTree-Gini)*	Gini	0.823
Metoda I (*CTree)*	średnia DiscPairs, Entropia, Gini	0.821
Metoda III (*VTree-Disc)*	DiscPairs	0.835
Metoda III (*VTree-Entropy)*	Entropia	0.837
Metoda III (*VTree-Gini)*	Gini	0.838
Metoda III (*VTree)*	średnia DiscPairs, Entropia, Gini	0.837

Tablica 9.38: Średnia dla 18 zbiorów danych dokładność predykcji.

duże znaczenie praktyczne, gdyż nie wymaga przeprowadzenia zabiegowego leczenia rewaskularyzacyjnego, a więc postępowania inwazyjnego. Celem identyfikacji "zdrowego serca" należy zdefiniować, co będzie rozumiane tutaj pod tym pojęciem.

Ze względu na zróżnicowane podejścia do wskazań do leczenia zabiegowego zwężeń tętnic wieńcowych, a w szczególności do granicznej wartości zwężenia określanego jako istotne (50% lub 70%) przyjęto 3 definicje. Pierwsza jako "zdrowe serce" przyjmuje stan, w którym brak jest jakichkolwiek zwężeń w jakichkolwiek naczyniach ('0-vessel disease'). W tym ujęciu stany z jakimkolwiek zwężeniem tętnic wieńcowych należeć będą do przeciwstawnego pojęcia, czyli tzw. "chorego serca". Definicja druga, do pojęcia 'zdrowe serce' przypisuje chorych, u których brak jest stenoz większych lub równych 50%. Według trzeciej definicji do stanu 'zdrowe serce' przynależeć będą pacjenci nie posiadający zwężeń tętnic wieńcowych większych lub równych 70%. Rozkłady klas decyzyjnych dla kolejno określonych definicji przedstawia Tab. 9.39. Zatem można powiedzieć, że są to dane zbalansowane.

'Zdrowe serce'	**Klasa 'TAK'**	**Klasa 'NIE'**
Def.1. Brak stenoz	79 (39.5%)	121 (60.5%)
Def.2. Brak zwężeń $\geq$ 50%	109 (54.5%)	91 (45.5%)
Def.3. Brak zwężeń $\geq$ 70%	118 (59%)	82 (41%)

Tablica 9.39: Liczebność klas decyzyjnych dla poszczególnych definicji pojęcia tzw. 'zdrowego serca' dla zbioru $HOLTER\_II$.

Wyniki testów przeprowadzonych celem identyfikacji 'zdrowego serca' dla wszystkich trzech definicji przedstawia Tab. 9.40. Jak wskazują wyniki, najlepszą dokładność klasyfikacji uzyskano dla definicji trzeciej, według której do poję-

'Zdrowe serce'	**Metoda I**	**Metoda II**	**Metoda III**	**Metoda IV**
Parametry metody	10-CV	10-CV	10-CV	10-CV, k=5
Def.1. Brak stenoz	0.654	0.606	0.557	0.772
Def.2. Brak zwężeń $\geq 50\%$	0.850	0.815	0.853	0.936
Def.2. Brak zwężeń $\geq 70\%$	0.939	0.924	0.949	0.966

Tablica 9.40: Wyniki (ACC) predykcji tzw. 'zdrowego serca' w CNS dla zbioru *HOLTER_II*.

cia 'zdrowe serce' należą obiekty bez zwężeń większych lub równych 70%. Słabsza efektywność dla definicji pierwszej może wiązać się z bardzo małymi, subtelnymi różnicami w zapisie EKG metodą Holtera między brakiem zwężeń a bardzo niewielkimi zwężeniami, które praktycznie nie zmieniają przepływu krwi w naczyniach.

Dodatkowy problem medyczny: 'Duże tętnice'

Kolejny problem badawczy polega na możliwości rozpoznawania zwężeń dużych tętnic wieńcowych, których zwężenia powodują zmiany o większym zasięgu niż takie same zmiany w mniejszych naczyniach. Wynika to z większego obszaru mięśnia sercowego zaopatrywanego przez duże naczynia. Jako duże przyjęto dla potrzeb tej rozprawy następujące tętnice wieńcowe (spośród wszystkich 10 dostępnych, ocenianych w koronarografii): lewa tętnica wieńcowa LCA (ang. *left coronary artery*), tętnica przednia zstępująca LAD (ang. *left anterior descending*), gałąź okalająca LCX (ang. *left circumflex artery*) oraz prawa tętnica wieńcowa RCA (ang. *right coronary artery*). LCA i RCA są głównymi naczyniami odchodzącymi od aorty, zaopatrującymi cały mięsień sercowy (LCA głównie lewy przedsionek, lewą komorę serca, 2/3 przedniej przegrody międzykomorowej, RCA głównie prawy przedsionek, prawą komorę i 1/3 tylną przegrody międzykomorowej). Naczynia LAD i LCX stanowią odgałęzienia LCA. Do pojęcia: zwężenia dużych tętnic wieńcowych będą przynależeć pacjenci, u których występuje co najmniej 70-cio % zwężenie co najmniej jednego ze wskazanych powyżej 'dużych' naczyń.

Rozkład klas decyzyjnych wynosił: 77 (38.5%) obiektów w klasie 'TAK' i 123 (61.5%) w klasie 'NIE'. Należy zwrócić uwagę, że przy tak zdefiniowanym problemie, do klasy decyzyjnej 'NIE' należą m.in. pacjenci z dużymi zwężeniami, ale mniejszych tętnic.

Celem oceny zdolności rozpoznawania istotnych ($\geq$70%) zwężeń dużych tętnic przeprowadzono badania na danych ze zbioru *HOLTER_II* z użyciem proponowanych metod. Rezultaty przedstawia Tab. 9.41, w której dla porównania umieszczono wyniki predykcji istotnych zwężeń dowolnych tętnic. Jak można za-

'Duże tt. wieńcowe'	**Metoda I**	**Metoda II**	**Metoda III**	**Metoda IV**
Parametry metody	10-CV	10-CV	10-CV	10-CV, k=5
Istotne zwężenia (≥70%)	0.834	0.856	0.902	0.866
Istotne zwężenia (≥70%) dużych tętnic (LCA, LAD, LCX, RCA)	0.856	0.799	0.831	0.844

Tablica 9.41: Wyniki (ACC) predykcji istotnych stenoz 'dużych tętnic' wieńcowych w CNS dla zbioru $HOLTER\_II$.

uważyć, klasyfikatory generalnie uzyskały podobne wyniki przy rozpoznawaniu istotnych zwężeń dużych tętnic, jak dla takich samych stenoz w dowolnych naczyniach, chociaż intuicyjnie oczekuje się, że zmiany dużych naczyń będą silniej wyrażone w danych i przez to lepiej rozpoznawane przez klasyfikatory. Jednym z możliwych wyjaśnień w tym przypadku może być fakt, że do badania wybierani byli pacjenci ze stabilną chorobą wieńcową, przyjmowani do planowej koronarografii. W stabilnej CNS o długotrwałym przebiegu może rozwijać się krążenie oboczne dla obszaru zaopatrywanego przez zwężone naczynie, kompensujące niedokrwienie tego obszaru.

Ogólnie, ograniczeniem proponowanych metod jest paradoksalnie konieczność udziału eksperta dziedzinowego w procesie tworzenia modelu, jednak korzyści jakie uzyskuje się dzięki zastosowaniu wiedzy dziedzinowej, takie jak poprawa dokładności, czułości, specyficzności, dodatniej i ujemnej wartości predykcyjnej klasyfikacji są warte takiej ceny.

Rozdział 10

Podsumowanie

Zawartość

W rozprawie przedstawiono propozycję metodyki stosowania wiedzy dziedzinowej do poprawiania jakości klasyfikatorów budowanych metodami drzewa decyzyjnego oraz k najbliższych sąsiadów. Cel główny rozprawy został sformułowany we wprowadzeniu (Rozdz. 1.2) w sposób następujący: *Opracowanie metod wykorzystujących wiedzę dziedzinową do poprawienia jakości klasyfikatorów tworzonych dwiema metodami, tj. metodą drzewa decyzyjnego oraz metodą k-NN.* Metody i algorytmy zaproponowane w rozprawie zostały zaimplementowane i przetestowane. Teza rozprawy została poddana weryfikacji empirycznej w oparciu o dane rzeczywiste. Porównano skuteczność proponowanych metod z wynikami klasyfikacji przeprowadzonej z użyciem innych znanych metod, takich jak naiwny klasyfikator Bayesa, drzewa i reguły decyzyjne, metoda wektorów nośnych SVM, metoda k-NN z klasycznymi miarami odległości, sztuczne sieci neuronowe oraz lasy losowe, opartej tylko na zbiorach danych.

Dodatkowo, poza weryfikacją otrzymanych klasyfikatorów za pomocą powszechnie znanych miar jakości klasyfikacji, tjakich jak np. dokładność, czułość czy swoistość, przeprowadzono szereg testów statystycznych, tj. test χ^2 i Wilcoxona. Testy, których celem było sprawdzenie istotności różnic między rzeczywistymi i przewidywanymi klasami decyzyjnymi, wskazały na zgodności wyników przewidywania z faktycznym rozkładem decyzji.

Wyniki eksperymentów przeprowadzone na danych dotyczących rzeczywistych złożonych problemów wykazują, że zaproponowane metody stosowania wiedzy dziedzinowej przewyższają efektywnością podejścia konwencjonalne oparte tylko na danych sensorowych. Potwierdzają tym założenia, że zaproponowane metody

zastosowania dodatkowej wiedzy dziedzinowej zwiększają jakość klasyfikatorów, udowadniając tym prawdziwość postawionych tez.

Do budowy modelu prognostycznego w rozpoznawaniu CNS wykorzystano dane kliniczne i EKG. Podejście takie może być bardzo przydatne dla klinicystów w prowadzeniu pacjentów z chorobą wieńcową, w szczególności wobec ograniczonego dostępu do inwazyjnego badania diagnostycznego, tj. koronarografii lub w przypadku występowania przeciwwskazań do jej wykonania (uczulenia na kontrast podawany podczas koronarografii, zły stan ogólny pacjenta, inne ostre schorzenia). U pacjentów z dodatnimi testami może być brany pod uwagę zabieg rewaskularyzacji, nawet jeśli inne wyniki badań wskazują na umiarkowane lub niskie ryzyko. W przypadku testów negatywnych, klinicysta może obserwować pacjenta kontynuując farmakoterapię.

Najbardziej atrakcyjnym z punktu widzenia medycznego aspektem tego podejścia jest zastosowanie łatwo dostępnych parametrów klinicznych, laboratoryjnych i elektrokardiograficznych, pozyskiwanych w sposób nieinwazyjny i tani. Szacowanie anatomii wieńcowej przed koronarografią może być przydatne przy podejmowaniu decyzji o interwencjach diagnostycznych i terapeutycznych.

Proponowane metody są ważne dla lekarzy, którzy leczą pacjentów z chorobą wieńcową w codziennej praktyce. Przewidywanie zwężenia tętnic wieńcowych może pomóc w lepszym prowadzeniu chorych i dostosowaniu leczenia CNS. Potrzebne są jednak dalsze badania w celu oceny, czy proponowane metody prowadzą do znaczących zmian w wynikach badań klinicznych oraz mogą być stosowane jako narzędzie wspomagające w procesie podejmowania decyzji klinicznych. W rozprawie podjęto temat ważny i aktualny, zarówno z poznawczego, jak i praktycznego punktu widzenia, wokół którego toczą się obecnie szerokie dyskusje.

Badania związane z rozprawą były wspierane przez grant Narodowego Centrum Nauki (projekt nr DEC-2013/09/B/ST6/01568) pod tytułem: *Wspomagane wiedzą dziedzinową wykrywanie z danych modeli procesów i ich zmian oraz zastosowanie opracowanych metod do przewidywania obecności istotnych zwężeń w tętnicach wieńcowych serca i powikłań interwencji zabiegowych*, realizowany na Uniwersytecie Rzeszowskim pod kierownictwem dr. hab. Jana Bazana, prof. UR oraz grant Narodowego Centrum Nauki (projekt nr DEC-2013/09/B/NZ5/00758) pod tytułem: *Rola receptorów kolagenowych, integryn alfa1beta1 i alfa2beta1, w powstawaniu reakcji zapalnej i zmian strukturalnych w drogach oddechowych w astmie - spojrzenie na drzewo oskrzelowe*, realizowany na Uniwersytecie Jagiellońskim pod kierownictwem dr n. med. Stanisławy Bazan-Socha. Planowane jest wykorzystanie zaproponowanych metod do analiz danych mikromacierzowych oraz ich wdrożenie w systemie dla Collegium Medicum Uniwersytetu Jagiellońskiego.

10.1 Wnioski i rezultaty

Na podstawie otrzymanych wyników można stwierdzić, że:

- Wszystkie zaproponowane metody wykorzystujące dodatkową wiedzę dziedzinową podczas konstrukcji klasyfikatorów uzyskują znacznie większą dokładność klasyfikacji niż podejścia konwencjonalne oparte tylko na zbiorach danych (bez dodatkowej wiedzy dziedzinowej). Efektywność klasyfikacji poprawia się o 20% przy zastosowaniu wzorców czasowych w drzewie decyzyjnym lokalnej dyskretyzacji do 35% w przypadku metody k-NN z odległością ontologiczną;

- Najkorzystniejsze wyniki klasyfikacji zwężeń tętnic wieńcowych w CNS uzyskuje się z zastosowaniem odległości semantycznej pomiędzy obiektami opartej na ontologii pojęć w metodzie k najbliższych sąsiadów;

- Spośród metod opartych na drzewie decyzyjnym największą poprawę skuteczności klasyfikacji otrzymano dla metody wykorzystującej cięcia weryfikujące do zwiększania wiarygodności podziałów węzłów drzewa;

- Metody klasyfikacji nadzorowanej z zastosowaniem dodatkowej wiedzy dziedzinowej podczas budowy drzew decyzyjnych i w metodzie k-najbliższych sąsiadów stanowią skuteczne algorytmy przewidywania obecności istotnych zwężeń tętnic wieńcowych w chorobie niedokrwiennej serca;

- Najlepszą skuteczność V-drzewa z cięciami weryfikującymi odnotowano dla zbiorów danych charakteryzujących się bardzo dużą liczbą atrybutów w stosunku do liczby obiektów;

- Wydajność V-drzew jest wysoka niezależnie od zastosowanej miary jakości cięć, co oznacza, że decydujące znaczenie ma zastosowanie dodatkowej wiedzy dziedzinowej zawartej w nadmiarowych cięciach;

- Algorytmy konstrukcji drzew decyzyjnych wykorzystujące dodatkową wiedzę dziedzinową ze względu na niską złożoność obliczeniową mogą być stosowane do klasyfikacji zbiorów danych z dużą liczbą cech (atrybutów);

- Metoda k najbliższych sąsiadów wykorzystująca odległość ontologiczną ze względu na dużą złożoność obliczeniową algorytmów opartych na odległościach, związaną z koniecznością wyliczenia odległości pomiędzy każdą parą obiektów może być stosowana do klasyfikacji na postawie małej liczby obiektów; na korzyść tej metody względem klasycznego podejścia wykorzystującego np. odległość Euklidesa przemawia znaczna redukcja liczby cech

wskazana przez ontologię pojęć definiowaną przez eksperta dziedzinowego (w medycznych zbiorach danych z 595 do 20); zastosowanie redukcji liczby atrybutów skraca czas obliczeń;

- Algorytm drzewa lokalnej dyskretyzacji wykorzystujący tylko wzorce czasowe jest najmniej skutecznym algorytmem klasyfikacji spośród wszystkich badanych wykorzystujących wiedzę dziedzinową;
- Weryfikacja zaproponowanych metod na większym medycznym zbiorze danych ($HOLTER\_II$) dała porównywalne lub lepsze wyniki klasyfikacji w odniesieniu do mniej licznego zbioru ($HOLTER\_I$), co potwierdza skuteczność zaproponowanych metod.

Podsumowując, w rozprawie zrealizowano następujące cele szczegółowe wykazując prawdziwość tezy:

- Opracowano metodę ekstrakcji cech opartą na tzw. wzorcach czasowych poprawiającą efektywność klasyfikatorów;
- Zaproponowano modyfikację oceny jakości podziału węzła przy generowaniu drzewa decyzyjnego zwiększającą skuteczność klasyfikacji za pomocą drzew;
- Wprowadzono metodę zwiększania wiarygodności podziałów węzłów drzewa decyzyjnego poprawiającą efektywność klasyfikacji z użyciem drzew;
- Zdefiniowano odległość semantyczną pomiędzy obiektami opartą na ontologii pojęć dającą zwiększenie wydajności klasyfikacji metodą k-NN;
- Przedstawiono opis wpływu czynnika modyfikującego percepcję testowanych obiektów w oparciu o modele klasyfikacji.

Przeprowadzone badania eksperymentalne pozwalają pozytywnie zweryfikować tezy rozprawy. Wykazano, że możliwe jest zaprojektowanie efektywnych klasyfikatorów wykorzystujących dodatkową wiedzę dziedzinową na etapie konstrukcji drzew decyzyjnych oraz definiowania odległości pomiędzy obiektami.

Przedstawione wyniki pokazują, że zastosowanie dodatkowej wiedzy dziedzinowej do budowy klasyfikatorów daje możliwość skutecznej diagnostyki choroby niedokrwiennej serca.

10.2 Kierunki dalszych badań

Przegląd tematyki dotyczącej rozprawy w zakresie perspektyw predykcji i opisu danych wskazuje, że problemy zastosowania dodatkowej wiedzy dziedzinowej w procesie klasyfikacji wymagają dalszych systematycznych badań. Należy podkreślić,

że pomimo wyraźnego rozwoju podejść uwzględniających dodatkową wiedzę dziedzinową, brak powszechnie obowiązującej i jednolitej metody reprezentacji tej wiedzy oraz brak jednoznacznie ustalonego etapu odkrywania wiedzy, na którym wiedzę dziedzinową powinno się stosować stanowią istotne ograniczenia w dokonywaniu wszelkich porównań i powszechnym wdrażaniu tej wiedzy w procesie KDD.

Przeprowadzone badania pokazują, że wykorzystanie odległości ontologicznej do wyznaczania podobieństwa obiektów w metodzie k-NN daje poprawę wydajność klasyfikacji. Niestety podczas obliczania odległości dla każdej pary obiektów występują znaczne ograniczenia czasowe i pamięciowe. Problem ten stanowi możliwy kierunek badań, który można rozwiązać poprzez obliczenia rozproszone z użyciem klastrów o rozproszonej pamięci.

Rozprawa oczywiście nie wyczerpuje całości problematyki. Można wskazać potrzebę badań dotyczących między innymi takich problemów jak:

- Grupowanie w oparciu o odległość semantyczną opartą na ontologii pojęć (np. metoda k-środków i hierarchiczna);
- Określenie mechanizmu weryfikacji jakości ontologii (np. spełnianie pewnych więzów ręcznie podanych przez eksperta);
- Aproksymacja odległości (jako decyzji) w oparciu o tablicę par obiektów treningowych, gdzie decyzja to odległość pomiędzy obiektami z pary;
- Potwierdzenie skuteczności klasyfikatorów wykorzystujących zaproponowane metody stosowania wiedzy dziedzinowej na dużej próbie standardową metodą, np. 'train and test';
- Modyfikacja miary jakości cięć weryfikujących w V-drzewie;
- Modyfikacja etapu algorytmu konstrukcji V-drzewa orzekającego o odłożeniu w czasie rozróżnienia obiektów, czyli w sytuacji gdy obiekty treningowe są kierowane do lewego i prawego poddrzewa, tylko gdy znacząca liczba podziałów weryfikujących inaczej je kieruje niż podział optymalny;
- Tworzenie specjalnego poddrzewa lub ogólniej: klasyfikatora (np. k-NN, SVM, dyskretyzacja hiperpłaszczyznowa) dla obiektów odłożonych podczas podziału węzła w V-drzewie;
- Opracowanie różnych metod rozstrzygania konfliktów podczas klasyfikacji za pomocą V-drzewa dla przypadku, gdy jedno z poddrzew zwróciło jedną wartość decyzji, a drugie poddrzewo inną;

- Opracowanie mechanizmu rezygnacji z podziału optymalnego w drzewach z cięciami weryfikującymi w sytuacji, gdy wybrany podział optymalny znacznie odbiega od podziału faktycznie optymalnego, na przykład dla nierepreprezentatywnego zbioru obiektów. W tym podejściu spośród k podziałów o zbliżonej jakości wybierane byłyby tylko te, które w sposób najbardziej podobny dzielą obiekty.

W rozprawie dokonano syntetycznego zestawienia wniosków wypływających z analiz, wykorzystując je do odniesienia ich do celów i postawionych tez.

Dodatek A

Dodatek medyczny

Zawartość

Stabilna choroba wieńcowa stanowi jedną z najczęstszych chorób układu sercowo-naczyniowego. Jest to zespół objawów klinicznych objawiający się występowaniem bólów w klatce piersiowej wywołanych wysiłkiem fizycznym, zimnem lub stresem. Podstawowym objawem, na który skarży się chory, jest uczucie ucisku, pieczenia bądź dławienia w okolicy zamostkowej, stąd choroba bywa nazywana także dławicą piersiową. Dolegliwości są wynikiem niedokrwienia mięśnia sercowego, którego przyczyną jest miażdżyca tętnic wieńcowych.

Choroba stanowi duży problem medyczny oraz społeczny. W większości krajów europejskich stwierdza się ją u 20000–40000 na milion mieszkańców. Ze względu na starzenie się populacji i coraz częstsze występowanie czynników ryzyka rozwoju choroby wieńcowej, takich jak: otyłość, cukrzyca typu 2 czy zespół metaboliczny obserwuje się stały wzrost częstości choroby wieńcowej na świecie.

W przeprowadzonych badaniach populacyjnych wykazano, że występowanie dławicy piersiowej wzrasta gwałtownie z wiekiem u obu płci, z 0,1–1% u kobiet w przedziale wiekowym 45–54 lat, do 10–15% w wieku 65–74 lat, natomiast u płci męskiej wzrasta z 2-5% do 11–20% w odpowiednich przedziałach wiekowych. Po 75 roku życia częstość ta jest zbliżona u obu płci [1].

A.1 Diagnostyka choroby wieńcowej

Diagnostyka choroby wieńcowej obejmuje badanie podmiotowe i przedmiotowe, badania laboratoryjne oraz specjalistyczne badania kardiologiczne, nieinwazyjne i inwazyjne. Podstawą rozpoznania jest prawidłowo zebrany wywiad, natomiast badania dodatkowe służą do potwierdzenia rozpoznania i oceny stopnia zaawansowania choroby. Najczęstszym i najbardziej typowym objawem zgłaszanym przez chorego jest ból w klatce piersiowej. Typowy ból związany z niedokrwieniem mięśnia sercowego to uczucie dyskomfortu w klatce piersiowej, ucisku, pieczenia czy dławienia. Zwykle zlokalizowany jest w okolicy zamostkowej i może promieniować do kończyn górnych, zwłaszcza lewej, pleców, nadbrzusza, szyi czy żuchwy.

Dolegliwości wywołują najczęściej takie czynniki jak: wysiłek fizyczny, zimne powietrze, stres, obfity posiłek oraz wzrost ciśnienia tętniczego. Ból trwa na ogół nie dłużej niż 10–15 minut i ustępuje po zaprzestaniu wysiłku lub zażyciu krótko działających nitratów, np. nitrogliceryny. Do nasilenia dolegliwości mogą prowadzić również: gorączka, anemia czy nadczynność tarczycy.

Charakterystyczne dla stabilnej choroby wieńcowej jest występowanie objawów, które nie nasilają się w okresie co najmniej 2 miesięcy. Stabilna dławica może wystąpić jako pierwsza manifestacja choroby niedokrwiennej serca, jak i u chorych po ostrym zespole wieńcowym (OZW).

Do najczęściej stosowanych skal, służących do oceny stopnia nasilenia dławicy, należy klasyfikacja zaproponowana przez Canadian Cardiovascular Society (CCS), przedstawiona w Tab. A.1.

Zgodnie z wytycznymi European Society of Cardiology (ESC), dotyczącymi postępowania w stabilnej chorobie wieńcowej, u każdego chorego należy oznaczyć profil lipidowy, poziom glukozy na czczo, morfologię oraz stężenie kreatyniny (klasa zaleceń I, poziom wiarygodności B, dla oznaczenia stężenia kreatyniny — C). Kontrolne badania, zarówno profilu lipidowego, jak i glikemii na czczo, powinny być wykonywane raz w roku (klasa zaleceń IIa, poziom wiarygodności C).

Badanie przedmiotowe pacjentów ze stabilną chorobą wieńcową w wielu przypadkach nie wykazuje odchyleń od normy, ale może ujawnić obecność czynników ryzyka miażdżycy tętnic. Ważne więc są pomiary ciśnienia tętniczego, wskaźnika masy ciała (BMI, ang. *Body Mass Index*) oraz obwodu talii.

A.1.1 Badania kardiologiczne nieinwazyjnie

Spośród nieinwazyjnych badań specjalistycznych w diagnostyce stabilnej choroby wieńcowej wykorzystuje się: elektrokardiogram spoczynkowy (EKG), elektrokardiograficzny test wysiłkowy, obrazowe próby obciążeniowe, echokardiografię spoczynkową (ECHO), 24-godzinne monitorowanie EKG metodą Holtera oraz tomografię komputerową.

Klasa	Opis
Klasa I	Zwyczajna codzienna aktywność fizyczna nie wywołuje dławicy; objawy występują przy większym i dłużej trwającym wysiłku fizycznym.
Klasa II	Niewielkie ograniczenie codziennej aktywności fizycznej; objawy pojawiają się: • przy szybkim chodzeniu po płaskim terenie lub szybkim wchodzeniu po schodach, po pokonaniu >200 m po terenie płaskim lub wejściu po schodach powyżej jednego piętra w normalnym tempie oraz przy chodzeniu po płaskim terenie lub po schodach po posiłkach, • przy wchodzeniu pod górę, • gdy jest zimno lub wieje wiatr, • pod wpływem stresu emocjonalnego lub w ciągu kilku godzin po przebudzeniu.
Klasa III	Znaczne ograniczenie codziennej aktywności fizycznej; objawy pojawiają się po przejściu 100–200 m po płaskim terenie bądź po wejściu po schodach na jedno piętro w normalnym tempie i w zwykłych warunkach.
Klasa IV	Każda aktywność fizyczna wywołuje dławicę piersiową; objawy mogą się również pojawiać w spoczynku.

Tablica A.1: Klasyfikacja dławicy piersiowej według CCS.

U każdego pacjenta z objawami sugerującymi obecność choroby wieńcowej, należy wykonać 12-odprowadzeniowy spoczynkowy elektrokardiogram, którego wartość diagnostyczna zwiększa się w przypadku występowania epizodu bólowego. Do najczęściej obserwowanych nieprawidłowości należą zmiany odcinka ST-T i obecność nieprawidłowego załamka Q, co może świadczyć o przebytym zawale serca, zaburzenia przewodnictwa przedsionkowo-komorowego oraz nadkomorowe lub komorowe zaburzenia rytmu. Jednak zmiany te nie są swoiste dla choroby wieńcowej i mogą być wywołane przerostem i przeciążeniem lewej komory serca, zaburzeniami elektrolitowymi czy stosowaniem leków antyarytmicznych. Należy jednak zwrócić uwagę, że prawie u połowy chorych ze stabilną dławicą piersiową nie występują zmiany w zapisie EKG [96].

Badaniem pierwszego wyboru u większości osób z podejrzeniem stabilnej cho-

roby wieńcowej jest elektrokardiograficzny test wysiłkowy. Jednak jego średnia czułość i swoistość w rozpoznawaniu choroby wieńcowej wynoszą jedynie 68 i 77% [76].

W diagnostyce choroby wieńcowej wykorzystuje się również obrazowe próby obciążeniowe, takie jak echokardiografia i scyntygrafia perfuzyjna z zastosowaniem obciążenia wysiłkiem fizycznym lub obciążenia farmakologicznego. Badania te charakteryzują się większą czułością i swoistością niż próba wysiłkowa w diagnozowaniu niedokrwienia (80–85% i 84–86% dla echokardiografii wysiłkowej oraz 85–90% i 70–75% dla scyntygrafii perfuzyjnej), pozwalają na lokalizację obszaru niedokrwienia oraz umożliwiają przeprowadzenie diagnostyki u chorych niezdolnych do wykonania wysiłku [41].

Badanie echokardiograficzne powinno być wykonywane u chorych z podejrzeniem wad zastawkowych, kardiomiopatii przerostowej lub niewydolności serca, a także po zawale serca oraz w przypadku obecności zmian w zapisie EKG, takich jak: blok lewej odnogi pęczka Hisa (LBBB, ang. *left bundle branch block*), blok przedniej wiązki lewej odnogi (LAH, ang. *left anterior hemiblock*) czy patologiczny załamek Q. U osób ze stabilną chorobą wieńcową, u których podejrzewa się zaburzenia rytmu, oraz w diagnostyce dławicy Prinzmetala zalecane jest 24-godzinne monitorowanie EKG metodą Holtera.

Wielorzędowa tomografia komputerowa jest metodą pozwalającą na nieinwazyjne obrazowanie tętnic wieńcowych, wykrywanie zwapnień oraz ocenę rozległości zmian. Ma jednak ograniczone znaczenie w diagnostyce zwężeń w naczyniach wieńcowych. W wytycznych ESC zaleca się wykonanie angio-CT tętnic wieńcowych u chorych z niejednoznacznym wynikiem elektrokardiograficznego testu wysiłkowego bądź obciążeniowego badania obrazowego, ale cechujących się niskim prawdopodobieństwem obecności choroby wieńcowej celem jej wykluczenia.

Badanie EKG metodą Holtera

W zapisach EKG metodą Holtera, czyli długotrwałych zapisach trwających co najmniej 24-godziny, ocenia się takie aspekty jak: rytm serca i jego zaburzenia (arytmie), zmiany odstępu PQ, zmiany odcinka ST czy zmienność rytmu serca HRV (ang. *heart rate variability*) oraz zmiany odstępu QT. Do ich oceny stosuje się m.in. takie parametry, jak wymienione poniżej.

Parametry rytmu serca

- Średnia, maksymalna i minimalna częstotliwość rytmu serca;
- Całkowita liczba analizowanych pobudzeń;
- Czas trwania zespołów QRS;

- Liczba dodatkowych pobudzeń pochodzenia komorowego (VE, ang. *ventricular extrasystolia*) i nadkomorowego (SVE, ang. *supraventricular extrasystolia*);
- Liczba epizodów tachykardii (minimalna częstotliwość rytmu na początku epizodu >125/min, a maksymalna na końcu 100/min);
- Liczba bradykardii (maksymalna częstotliwość rytmu na początku epizodu <45/min, a minimalna na końcu >55/min);
- Liczba pauz (brak zespołów QRS w okresie dłuższym niż 2380 ms po zespole komorowym, 1800 ms po zespole nadkomorowym, 2000 ms po zespole dominującym).

Zmienność rytmu serca HRV opisuje różnice w długościach interwałów RR wyznaczanych przez kolejne szczyty zespołów QRS. Występowanie tych różnic świadczy o zdolności serca do adaptacji względem bodźców zewnętrznych i informuje o pracy autonomicznego systemu nerwowego.

Do oceny dobowej zmienności rytmu serca wykorzystuje się parametry analizy czasowej (badanie czasu trwania kolejnych odstępów NN, czyli odstępów między pobudzeniami zatokowymi oraz różnic między nimi) oraz częstotliwościowej (badanie widma cyklicznie występujących zmian czasu trwania kolejnych odstępów NN) w wybranych przedziałach czasu. Analizę w dziedzinie częstotliwości zwykle przeprowadza się metodą tzw. szybkiej transformacji Fouriera (ang. *fast Fourier transformation*).

Parametry HRV

- SDRR (ang. *standard deviation of RR*)- odchylenie standardowe długości odstępów RR dla wszystkich uderzeń, gdzie RR to odstęp pomiędzy kolejnymi załamkami R zapisu EKG;
- SDNN (ang. *standard deviation of NN*) - odchylenie standardowe czasów trwania wszystkich odstępów NN, gdzie NN (ang. *normal–normal*) to odstęp pomiędzy kolejnymi pobudzeniami zatokowymi, N oznacza pobudzenie prawidłowe, w badanym okresie (w ciągu doby lub poszczególnych godzin), oznaczane w [ms];
- TINN - trójkątna interpolacja odstępów NN; długość podstawy w [ms] trójkąta aproksymującego histogram kolejnych odstępów RR rytmu zatokowego;
- SDANN (ang. *standard deviation of averaged NN intervals*) - odchylenie standardowe średnich wartości odstępów NN mierzonych w kolejnych krótkich przedziałach (5-minutowych) dłuższego okresu (całej doby lub poszczególnych godzin);

- SDNNI (SDNN index) - wskaźnik SDNN; średnia z odchyleń standardowych kolejnych odstępów NN z kolejnych 5-minutowych okresów badania, oznaczany w [ms], z danego przedziału czasu;
- RMSSD (ang. *root mean square of successive differences*) – pierwiastek kwadratowy średniej z sumy kwadratów różnic pomiędzy kolejnymi odstępami NN w badanym okresie;
- NN_50 - ilość par przylegających interwałów NN, które różnią się o więcej niż 50 milisekund (tylko w rytmie zatokowym);
- pNN50 (ang. *perecentage of NN intervals*) – odsetek odstępów NN różniących się od sąsiednich o ponad 50 ms względem liczby wszystkich odstępów NN w badanym okresie;
- ULF (ang. *ultra low frequency*) - pasmo w widmie zmienności rytmu serca o najniższej częstotliwości, tj. <0.0033 Hz;
- VLF (ang. *very low frequency*) - pasmo w widmie zmienności rytmu serca o bardzo niskiej częstotliwości tj. 0.0033-0.04 Hz;
- LF (ang. *low frequency*) – pasmo w widmie zmienności rytmu serca o niskiej częstotliwości (0.04 – 0.15 Hz). Norma dla 5 min: 1170 +/- 416 $[ms^2]$;
- HF (ang. *high frequency*) – pasmo w widmie zmienności rytmu serca o wysokiej częstotliwości (0.15 – 0.4 Hz). Norma dla 5 min: 975 +/- 203 $[ms^2]$;
- Stosunek LF do HF. Norma dla 5 min: 1.5-2.0.

Parametry odstępu PQ

- Minimalny, maksymalny, średni czas trwania odstępu PQ z 30 sekund;
- Czas trwania odstępów PQ.

Parametry odcinka ST

- Obniżenia i uniesienia odcinków ST;
- Liczba epizodów ST.

Parametry odstępu QT

- Minimalny, maksymalny, średni czas trwania odstępu QT z 30 sekund;
- Minimalny, maksymalny, średni czas trwania skorygowanego odstępu QT z 30 sekund (korekcja np. Bazzeta);
- Dyspersja QT.

A.1.2 Badania kardiologiczne inwazyjnie

Do inwazyjnych badań diagnostycznych zalicza się koronarografię tętnic wieńcowych, która jest uznawana za podstawową metodę diagnostyki choroby wieńcowej, umożliwiającą ocenę anatomii tętnic wieńcowych oraz lokalizację zwężeń. Koronarografia pozwala na ustalenie sposobu postępowania (udrażnianie i/lub farmakoterapia) i określenie rokowania. Ze względu na inwazyjny charakter jest jednak obarczona ryzykiem wystąpienia objawów ubocznych. Ryzyko poważnych powikłań związanych z koronarografią wynosi 1–2%, natomiast częstość zgonów, zawałów serca lub udarów mózgu łącznie określa się na 0.1-0.2% [41].

A.2 Postępowanie w stabilnej chorobie wieńcowej

Celem leczenia choroby wieńcowej jest poprawa jakości życia oraz rokowania. Istnieją dwa podstawowe sposoby leczenia tej choroby: farmakoterapia oraz udrażnianie, czyli rewaskularyzacja — przezskórna lub chirurgiczna. Ważna jest również modyfikacja odwracalnych czynników ryzyka, takich jak: zmniejszenie masy ciała, zaprzestanie palenia tytoniu czy regularna aktywność fizyczna. Poza tym należy dążyć do optymalnej terapii chorób współistniejących, tj. nadciśnienia tętniczego i cukrzycy.

A.2.1 Farmakoterapia

Do leków o udokumentowanym działaniu, prowadzącym do zmniejszenia śmiertelności, a przez to poprawy rokowania u osób ze stabilną chorobą wieńcową, należą: kwas acetylosalicylowy (ASA, ang. *acetylsalicylic acid*), statyny, β-adrenolityki oraz inhibitory konwertazy angiotensyny (ACE, ang. *angiotensin converting enzyme*).

Do leków zwalczających objawy choroby wieńcowej należą nitraty, β - adrenolityki oraz antagoniści wapnia. Leki te łagodzą objawy dławicy piersiowej poprzez zmniejszanie zapotrzebowania mięśnia sercowego na tlen lub zwiększenie przepływu krwi do niedokrwionych obszarów serca.

A.2.2 Udrażnianie tętnic wieńcowych

Rewaskularyzację należy rozważyć, gdy dotychczasowe leczenie okazuje się nieskuteczne. Jej celem jest ograniczenie objawów oraz przedłużenie życia pacjenta. Istnieją dwa ugruntowane podejście do udrażniania: rewaskularyzacja chirurgiczna CABG (ang. *Coronary artery bypass graft*) i przezskórna interwencja wieńcowa PCI (ang. *percutaneous coronary intervention*) [1]. PCI może oznaczać angioplastykę balonową, określaną też PTCA (ang. *Percutaneous Transluminal Coronary*

Angioplasty) z lub bez implantacji stentów lub aterektomię (wycięcie blaszki miażdżycowej).

W badaniu *Asymptomatic Cardiac Ischaemia Pilot Study (ACIP)* porównywano skuteczność udrażniania przezskórnego lub chirurgicznego z farmakoterapią u osób ze stabilną chorobą wieńcową i bezobjawowym niedokrwieniem, udokumentowanym w próbie wysiłkowej lub ambulatoryjnym EKG. Na podstawie wyników stwierdzono, że chorzy z grupy wysokiego ryzyka odnoszą istotnie większe korzyści z rewaskularyzacji serca niż z farmakoterapii [44]. W metaanalizie obejmującej badania porównawcze rewaskularyzacji chirurgicznej z farmakoterapią, oraz w innych badaniach obserwacyjnych wykazano, że operacyjne leczenie stabilnej choroby wieńcowej wpływa na poprawę rokowania jedynie u osób z grup umiarkowanego i wysokiego ryzyka [172].

U pozostałych chorych natomiast należy rozważać udrażnianie przezskórne lub leczenie farmakologiczne. Dowiedziono, że rewaskularyzacja przezskórna w porównaniu z farmakoterapią nie prowadzi do zmniejszenia umieralności pacjentów ze stabilną chorobą wieńcową, jest natomiast skuteczniejsza w zakresie redukcji liczby incydentów sercowo-naczyniowych, wpływających na jakość życia chorych [32]. Badanie*Angioplasty Compared to Medicine (ACME)* wykazało istotną przewagę leczenia inwazyjnego nad leczeniem zachowawczym choroby wieńcowej. U pacjentów poddanych zabiegowi przezskórnej rewaskularyzacji stwierdzono lepszą wydolność fizyczną i rzadziej obserwowano objawy, natomiast częstość zgonów i zawałów serca była porównywalna w obu grupach [54]. Do podobnych wniosków prowadzi badanie *Second Randomised Intervention Treatment of Angina (RITA-2)*, w którym okazało się, że rewaskularyzacja przezskórna, lepiej niż farmakoterapia wpływa na zmniejszenie objawów niedokrwienia i poprawę wydolności fizycznej, jednak wiązała się częściej ze śmiercią i zawałem serca w okresie okołozabiegowym [68].

Wybór postępowania terapeutycznego wobec szerokich możliwości leczenia stabilnej choroby wieńcowej, powinien być uzależniony od wielu czynników, przede wszystkim od korzyści odniesionych przez pacjenta z zaproponowanego sposobu leczenia, przy jak najmniejszym ryzyku powikłań, możliwości zapewnienia pełnej rewaskularyzacji, ograniczenia chorób współistniejących oraz preferencji i zgody pacjenta.

Dodatek B

Dodatek dotyczący hurtowni danych medycznych

Zawartość

Dla potrzeb importowania, przechowywania i przetwarzania danych pochodzących z rejestratorów zapisu EKG metodą Holtera (system HolCARD EKG firmy Aspel w przypadku zbioru $HOLTER\_I$ [9] i R12 systemu BTL Holter H600 dla zbioru $HOLTER\_II$ [31]) zaprojektowano i zaimplementowano hurtownie danych medycznych. Jako główne założenia przyjęto:

- Hurtownia powinna umożliwiać przechowywanie dużej ilości danych medycznych;
- Hurtownia powinna przechować wszystkie dane, w tym kliniczne oraz dane zapisu EKG metodą Holtera;
- Dostęp do danych hurtowni powinien być szybki i niezawodny;
- Dostęp powinien być autoryzowany ze względu na wrażliwość danych medycznych.

Dane pacjentów pozyskano i zapisano na nośniku elektronicznym w postaci plików binarnych (EKG Holter) oraz tekstowych (dane kliniczne). W pierwszym

etapie zweryfikowano kompletność danych pacjentów, takich jak daty badań, wyniki. W kolejnym etapie dane EKG zostały wczytane oraz zagregowane przy użyciu specjalistycznego oprogramowania (system HolCARD 24W EKG firmy Aspel w przypadku zbioru *HOLTER_I* i CardioPoint-Holter w wersji v2-23 dla zbioru *HOLTER_II*). Następnie dane pacjentów oraz dane zagregowane z systemów HolCARD oraz CardioPoint-Holter zostały zaimportowane do bazy przy pomocy zaimplementowanego w środowisku Java dedykowanego importera. Baza danych została umieszczona na serwerze będącym częścią Interdyscyplinarnego Centrum Modelowania Komputerowego (ICMK) Uniwersytetu Rzeszowskiego.

B.1 System zarządzania relacyjną bazą danych

Do projektowania i implementacji hurtowni danych wykorzystano system zarządzania relacyjną bazą danych PostgreSQL. Zastosowanie tego silnika oraz zaprojektowanie relacyjnej bazy danych pozwoliło na osiągnięcie wszystkich założeń. Do przechowania danych medycznych zostały zaprojektowane i utworzone dwie hurtownie danych ze względu na odmienne formaty i zawartości plików eksportowanych przez obydwa wykorzystywane systemy Holtera (Aspel i BTL).

B.1.1 Zbiór *HOLTER_I*

Do przechowania danych zbioru *HOLTER_I* została utworzona hurtownia danych składająca się z 10 tabel, połączonych ze sobą wzajemnymi relacjami. Listę tabel przedstawia kod B.1. Opis zawartości każdej z tabel przedstawia Tab. B.1.

B.1: Tabele w zbiorze HOLTER_I

```
+-----------------------+
| Tables                |
+-----------------------+
| clinics               |
| exams                 |
| hrv                   |
| patients              |
| patients_names        |
| qt                    |
| qthour                |
| signals               |
| st                    |
| sve                   |
+-----------------------+
```

Tabela	**Zawartość**
clinics	dane kliniczne pacjentów
exams	dane wykonanych badań Holtera
hrv	dane dotyczące dobowej zmienności rytmu serca HRV
patients	dane pacjentów (bez danych "wrażliwych")
patients_names	dane "wrażliwe" pacjentów
qt	dane dotyczące odstępu QT (dane co 5 minut)
qthour	dane dotyczące odstępu QT (dane godzinowe)
signals	surowy sygnał
st	dane dotyczące odcinka ST
sve	dane na temat rytmu serca i jego zaburzeń

Tablica B.1: Tabele bazy danych zbioru *HOLTER_I*.

Fragment tabeli *clinics* przedstawia kod B.2.

B.2: Tabela *clinics* w zbiorze HOLTER_I

Field	Type	**Null**	**Key**	**Default**	Extra
ID_CLINIC	bigint(20)	YES		**NULL**	
ID_PATIENT	bigint(20)	YES		**NULL**	
EX_DATE	**varchar**(20)	YES		**NULL**	
ZYFLO	**smallint**(6)	YES		**NULL**	
AGE	**smallint**(6)	YES		**NULL**	
PTCA	**smallint**(6)	YES		**NULL**	
ANGIO_OUT	**varchar**(500)	YES		**NULL**	
CABG	tinyint(4)	YES		**NULL**	
ATHEROM	tinyint(4)	YES		**NULL**	
CRP	**float**	YES		**NULL**	
FIBRYN	**float**	YES		**NULL**	
TSH	**float**	YES		**NULL**	
D_DIM	**float**	YES		**NULL**	
TROP_1	**varchar**(10)	YES		**NULL**	
TROP_2	**varchar**(10)	YES		**NULL**	

Fragment tabeli *exams* przedstawia kod B.3.

B.3: Tabela *exams* w zbiorze HOLTER_I

Field	Type	**Null**	**Key**	**Default**	Extra

ID_HOLTER	bigint(20)	YES		**NULL**	
ID_CLINIC	bigint(20)	YES		**NULL**	
START	**varchar**(30)	YES		**NULL**	
END	**varchar**(30)	YES		**NULL**	
TIME	**varchar**(10)	YES		**NULL**	

Tabelę *st* przedstawia kod B.4, a jej przykładowe dane kod B.5

B.4: Tabela *st* w zbiorze HOLTER_I

Field	Type	**Null**	**Key**	**Default**	Extra
ID_HOLTER	bigint(20)	YES		**NULL**	
HOUR	tinyint(4)	YES		**NULL**	
TIME	**varchar**(20)	YES		**NULL**	
MAXHR_DOM	**int**(11)	YES		**NULL**	
MINHR_DOM	**int**(11)	YES		**NULL**	
AVGHR_DOM	**int**(11)	YES		**NULL**	
QRS_ST_DOM	**int**(11)	YES		**NULL**	
ARTEF_DOM	**float**	YES		**NULL**	
ST_UP1	**float**	YES		**NULL**	
ST_UP2	**float**	YES		**NULL**	
ST_UP3	**float**	YES		**NULL**	
ST_DOWN1	**float**	YES		**NULL**	
ST_DOWN2	**float**	YES		**NULL**	
ST_DOWN3	**float**	YES		**NULL**	
ST_EPI	**int**(11)	YES		**NULL**	

B.5: Przykładowe dane tabeli *st* w HOLTER_I

ID_HOLTER	**HOUR**	**TIME**	MAXHR_DOM	MINHR_DOM	AVGHR_DO
1301843768731	0	Razem:	96	48	72
1301843768731	1	12:55	90	67	79
1301843768731	2	13:55	92	65	79

B.1.2 Zbiór *HOLTER_II*

Do przechowania danych zbioru *HOLTER_II* utworzona została hurtownia danych składająca się z 15 tabel, połączonych ze sobą wzajemnymi relacjami. Listę

tabel przedstawia kod B.6, a krótki opis każdej z tabel Tab. B.2.

B.6: Tabele w zbiorze HOLTER_II

```
+-------------------+
| Tables            |
+-------------------+
| apnea             |
| beats             |
| clinics           |
| edf_annotations   |
| exams             |
| hrv               |
| patients          |
| patients_names    |
| pmi               |
| pq                |
| qt                |
| rhythm            |
| signals           |
| st                |
| sve               |
+-------------------+
```

Tabela apnea

Głównym zadaniem tabeli jest przechowywanie informacji dotyczących bezdechów, zaimportowanych z pliku Apnea*.csv pacjentów. Zawiera 9 pól do przechowywania danych. Narzędzia dotyczące bezdechu zostały umieszczone przez producenta eksperymentalne i służą jedynie do celów testowych. Obecnie niewykorzystywane przez system.

Tabela beats

Głównym zadaniem tabeli jest przechowywanie informacji dotyczących załamków EKG, zaimportowanych z pliku BeatTable*.csv pacjentów. Zawiera 9 pól do przechowywania danych: id_holter – unikalny identyfikator badania Holtera, time – czas wystąpienia załamka R, type – Klasa uderzeń, gdzie: N: Prawidłowe uderzenie, S: Przedwczesne pobudzenie nadkomorowe, V: Przedwczesne pobudzenie komorowe, B: Blok odnogi pęczka Hisa lub zaburzenia przewodnictwa śródkomorowego, Q: Uderzenie niemożliwe do określenia (wątpliwe), X: Artefakt (w założeniu, nie ma uderzenia), rr – czas trwania odstępu RR , p_on – odległość początku załamka

Tabela	**Zawartość**
apnea	informacje dotyczące bezdechów
beats	informacje dotyczące załamków EKG
clinics	dane kliniczne pacjentów
edf_annotations	adnotacje dotyczące surowego zapisu sygnału
exams	dane badań Holtera pacjentów
hrv	dane dotyczące dobowej zmienności rytmu serca HRV
patients	dane identyfikacyjne pacjentów
patients_names	dane "wrażliwe" pacjentów
pmi	dane na temat pracy rozrusznika serca
pq	dane dotyczące odstępu PQ
qt	dane dotyczące odstępu QT
rhythm	dane na temat rytmu serca
signals	surowy sygnał
st	dane dotyczące odcinka ST
sve	dane na temat arytmii

Tablica B.2: Tabele bazy danych zbioru *HOLTER_II*.

P od załamka R, p_off – odległość końca załamka P od załamka R, qrs_on – odległość początku zespołu QRS od załamka R, qrs_off – odległość końca zespołu QRS od załamka R, t_off – odległość końca załamka T od załamka R.

Tabela clinics

Głównym zadaniem tabeli jest przechowywanie danych klinicznych pacjentów pochodzących z dokumentacji medycznej. Zawiera 143 pola do przechowywania danych takich jak: typ choroby niedokrwiennej serca CNS (stabilna/niestabilna), klasyfikacja CNS wg Canadian Cardiovascular Society (klasy: I-IV), obecność nadciśnienia tętniczego HA (Hypertonia Arterialis), stopień nasilenia HA, obecność cukrzycy DM (Diabetes Mellitus), typ cukrzycy, podwyższony poziom lipidów, stosowanie używek (papierosy), obecność przewlekłej niewydolności serca ICC (Insufficientia Circulatoria Chronica), klasyfikacja ciężkości objawów ICC wg New York Heart Association (klasy: I-IV), frakcja wyrzutowa EF (Ejection Fraction), obecność: miażdżycy zarostowej tętnic kończyn dolnych PAOD (Peripheral Arterial Occlusive Disease), skala Rutherford (kategorie: 0-6), obecność: POCHP (przewlekła obturacyjna choroba płuc), otyłości, przewlekłej niewydolności nerek (Insufficientia Renalis Chronica), waga pacjenta, wzrost, bmi pacjenta, liczba krwinek białych, liczba krwinek czerwonych, hemoglobina, hematokryt, liczba płytek krwi.

Tabela edf_annotations

Zadaniem tabeli jest przechowywanie adnotacji zaimportowanych z pliku *.edf pacjentów, w którym zapisane są informacje o surowym sygnale. Zawiera 5 pól do przechowywania danych.

Tabela exams

Zadaniem tabeli jest przechowywanie danych badań Holtera pacjentów. Zawiera 5 pól do przechowywania danych: unikalny identyfikator badania Holtera, identyfikator badania klinicznego, czas rozpoczęcia badania, czas zakończenia badania i data badania.

Tabela hrv

Zadaniem tabeli jest przechowywanie danych dotyczących dobowej zmienności rytmu serca (HRV – heart rate variability), zaimportowanych z pliku Overview*.csv. Zawiera 22 pola do przechowywania danych, takie jak: sdrr - odchylenie standardowe długości odstępów RR dla wszystkich uderzeń, sdnn (standard deviation of NN) - odchylenie standardowe czasów trwania wszystkich odstępów NN czy lf (low frequency) – pasmo w widmie zmienności rytmu serca o niskiej częstotliwości (0.04 – 0.15 Hz).

Tabela patients

Zadaniem tabeli jest przechowywanie danych identyfikacyjnych pacjentów (bez danych "wrażliwych"). Zawiera 3 pola do przechowywania danych.

Tabela patients_names

Zadaniem tabeli jest przechowywanie danych "wrażliwych" pacjentów. Zawiera 4 pola do przechowywania danych.

Tabela pmi

Zadaniem tabeli jest przechowywanie danych na temat pracy rozrusznika serca (ang. *pacemaker implantation*) zaimportowanych z pliku Overview*.csv. Zawiera 11 pól do przechowywania danych. Niewykorzystana w badaniach ze względu na niekwalifikowanie do analiz pacjentów z rozrusznikami.

Tabela pq

Zadaniem tabeli jest przechowywanie danych dotyczących odstępu PQ, zaimportowanych z pliku Overview*.csv Zawiera 7 pól do przechowywania danych, takich jak: minimalny czas trwania odstępu PQ z 30 sekund, średni czas trwania odstępu PQ z 30 sekund czy maksymalny czas trwania odstępu PQ z 30 sekund.

Tabela qt

Zadaniem tabeli jest przechowywanie danych dotyczących odstępu QT, zaimportowanych z pliku Overview*.csv Zawiera 10 pól do przechowywania danych, takich jak: minimalny czas trwania odstępu QT z 30 sekund, średni czas trwania odstępu QT z 30 sekund czy minimalny czas trwania skorygowanego odstępu QT z 30 sekund (korekcja Bazzeta).

Tabela rythm

Zadaniem tabeli jest przechowywanie danych na temat rytmu serca, zaimportowanych z pliku Overview*.csv Zawiera 20 pól do przechowywania danych, takich jak: najniższa, średnia i najwyższa wartość rytmu czy czas trwania tachykardii.

Tabela signals

Zadaniem tabeli jest przechowywanie danych, zaimportowanych z pliku *.edf (surowy sygnał). Zawiera 14 pól do przechowywania danych, w tym na 12 odprowadzeń klasycznego elektrokardiogramu: 3 kończynowe dwubiegunowe Einthovena: I, II, III, 3 kończynowe jednobiegunowe Goldbergera: aVR, aVL, aVF, 6 przedsercowych jednobiegunowych: V1-6 (albo C1-6).

Tabela st

Zadaniem tabeli jest przechowywanie danych dotyczących odcinka ST, zaimportowanych z pliku Overview*.csv Zawiera 28 pól do przechowywania danych, takich jak: maksymalne uniesienia odcinka ST w odprowadzeniu dwubiegunowym kończynowym I czy maksymalne obniżenia odcinka ST w odprowadzeniu dwubiegunowym kończynowym I.

Tabela sve

Zadaniem tabeli jest przechowywanie danych na temat pobudzeń dodatkowych (V - komorowych, ventricular, S - nadkomorowych, supraventricular), zaimportowanych z pliku Overview*.csv. Zawiera 22 pola do przechowywania danych, takie jak:

liczba pojedynczych pobudzeń komorowych, liczba bigeminii komorowych, liczba pojedynczych pobudzeń nadkomorowych czy liczba przerw.

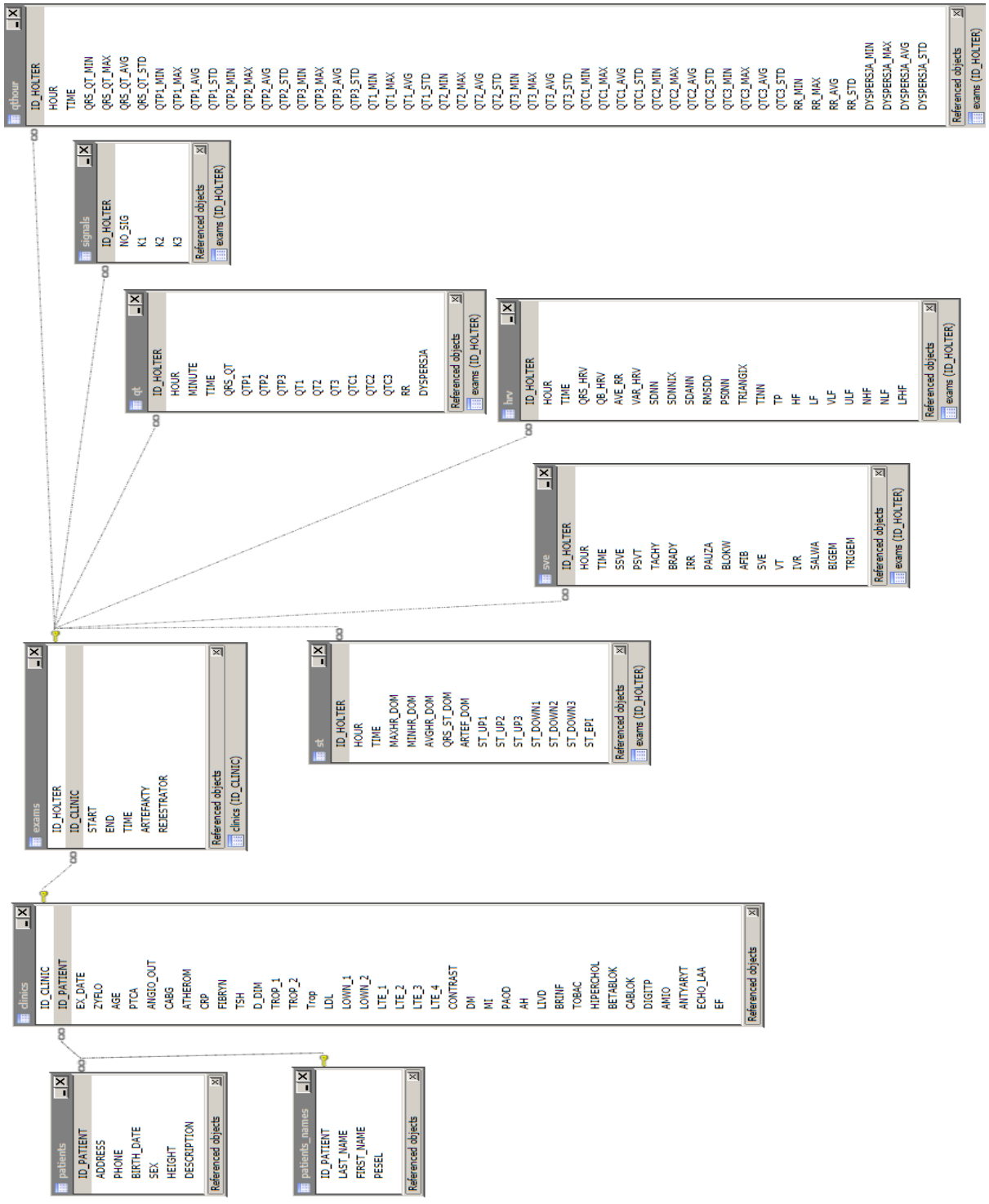
B.2 Relacje w bazie danych

W bazie danych, dzięki zaprojektowanej strukturze, nie występuje redundancja (nadmiarowość, powtarzanie się tych samych informacji) danych. Relacje w zaprojektowanej bazie danych spełniają założenia pierwszej postaci normalnej:

- Opisują jeden obiekt;
- Wartości atrybutów są elementarne – każda kolumna jest wartością skalarną (atomową);
- Nie zawiera kolekcji;
- Posiada klucz główny;
- Kolejność wierszy może być dowolna (znaczenie danych nie zależy od kolejności wierszy).

B.3 Diagram ERD (diagram związków encji)

Zaprojektowana baza danych spełnia wszystkie wymagania odnośnie funkcjonalności i pozwala na przechowywanie wszystkich danych medycznych pozyskanych do badań eksperymentalnych. Zaproponowany schemat pozwala na przeprowadzenie wszystkich założonych badań naukowych z wykorzystaniem pozyskanych danych medycznych. Schemat bazy danych dla zbioru $HOLTER\_I$ przedstawia Rys. B.1.



Rysunek B.1: Schemat ERD bazy danych zbioru *HOLTER_I*.

B.4 Przykładowe zapytania

Przykładowe zapytanie SQL do bazy wypisujące dane kliniczne oraz informacje na temat wykonanych badań diagnostycznych pacjenta o identyfikatorze 50 przedstawia kod B.7.

B.7: Przykładowe zapytanie SQL.

```
SELECT *
FROM CLINICS C, PATIENTS P
WHERE P.ID_PATIENT=50 AND P.ID_PATIENT=C.ID_PATIENT
```

Przykładowe zapytanie SQL wyliczające liczbę pacjentów z wykonanym zabiegiem wszczepienia pomostu naczyniowego (CABG) w lewej tętnicy wieńcowej przedstawia kod B.8.

B.8: Przykładowe zapytanie SQL.

```
SELECT *
COUNT(CLINICS.ID_PATIENT)
FROM CLINICS
WHERE PAST_LCA_BYPASS='TAK'
```

Konstrukcja hurtowni umożliwia wczytanie dowolnej liczby badań klinicznych oraz zapisów Holtera dla każdego pacjenta.

Spis rysunków

Spis tablic

Spis algorytmów

Ważniejsze oznaczenia i skróty

Symbol	Opis	Strona
CNS	Choroba niedokrwienna serca	8
PCI	Przezskórna interwencja wieńcowa (Percutaneous Coronary Intervention)	8
CABG	Pomostowanie tętnic wieńcowych (Coronary Artery Bypass Graft)	8
DM	Eksploracja danych (Data Mining)	16
KDD	Odkrywanie wiedzy (Knowledge Discovery)	16
WD	Wiedza dziedzinowa (domain knowledge)	16
$\mathbf{SI} = (U, A)$	System informacyjny (information system)	23
DT	Tablica decyzyjna (decision table)	24
CTree	Klasyczne drzewo lokalnej dyskretyzacji	55
$Q_{Disc}(c, X)$	Miara jakości podziałów w drzewie *CTree* oparta na liczbie par obiektów rozróżnianych przez cięcie	55
$Q_{Entropy}(c, X)$	Miara jakości podziałów w drzewie *CTree* oparta na entropii	55
$Q_{Gini}(c, X)$	Miara jakości podziałów w drzewie *CTree* oparta na indeksie Giniego	56
$c = (a, v)$	Cięcie wyznaczające podział obiektów	53
T lub $T(c)$	Wzorzec definiowany przez cięcie c	53
$TL(c)$	Wzorzec lewy definiowany przez cięcie c	53
$TR(c)$	Wzorzec prawy definiowany przez cięcie c	53
$\mathbf{A}(T)$	Zbiór obiektów z U pasujących do wzorca T	54
CTree-Disc	Klasyfikator zbudowany za pomocą *CTree* i miary Q_{Disc}	58
CTree-Entropy	Klasyfikator zbudowany za pomocą *CTree* i miary $Q_{Entropy}$	58
CTree-Gini	Klasyfikator zbudowany za pomocą *CTree* i miary Q_{Gini}	58
TN	Liczba prawidłowych klasyfikacji przykładów negatywnych (True Negatives)	61
FP	Liczba nieprawidłowych klasyfikacji przykładów negatywnych (False Positives)	61
FN	Liczba nieprawidłowych klasyfikacji przykładów pozytywnych (False Negatives)	61
TP	Liczba prawidłowych klasyfikacji przykładów pozytywnych(True Positives)	61

Bibliografia

[1] Guidelines on the management of stable angina pectoris: executive summary: the Task Force on the Management of Stable Angina Pectoris of the European Society of Cardiology. European Heart Journal 27 (2006) 1341–1381.

[2] Y. S. Abu-Mostafa. Hints. Neural Computation 7 (1995) 639–671.

[3] S. M. Al-Khatib, C. W. Yancy, P. Solis, L. Becker, E. J. Benjamin, R. G. Carrillo, J. A. Ezekowitz, G. C. Fonarow, B. K. Kantharia, M. Kleinman, G. Nichol, P. D. Varosy. 2016 AHA/ACC clinical performance and quality measures for prevention of sudden cardiac death: A report of the American College of Cardiology/American Heart Association Task Force on Performance Measures. Circulation: Cardiovascular Quality and Outcomes 10(2) (2017) e000022.

[4] A. Alizadeh, M. Eisen, R. Davis, C. Ma, et. al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403 (2000) 503–511.

[5] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levinei. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96(12) (1999) 6745–6750.

[6] D. G. Altman. Practical Statistics for Medical Research. Chapman and Hall/CRC, London, 1997.

[7] R. Ambrosino, B. Buchanan. The use of physician domain knowledge to improve the learning of rule-based models for decision-support. In: Proceedings of the Annual Fall Symposium of the American Medical Informatics Association. Washington DC., 1999, pp. 192–196.

[8] S. S. Anand, D. A. Bell, J. G. Hughes. The role of domain knowledge in data mining. In: Proceedings of the fourth international conference on information

and knowledge management. ACM Press, New York, NY, USA, 1995, pp. 37–43.

[9] ASPEL S.A. Strona internetowa: `http://www.aspel.com.pl/`.

[10] Y. Bar-Yam. Dynamics of Complex Systems. Addison Wesley, New York, USA, 1997.

[11] J. Bazan, S. Bazan-Socha, S. Buregwa-Czuma, L. Dydo, W. Rzasa, A. Skowron. A classifier based on a decision tree with verifying cuts. In: Proceedings of the Workshop on CS&P 2014, Chemnitz, Germany. 2014, Informatik-Bericht, Humboldt University, vol. 245, pp. 13–21.

[12] J. Bazan, S. Bazan-Socha, S. Buregwa-Czuma, L. Dydo, W. Rzasa, A. Skowron. A classifier based on a decision tree with verifying cuts. Fundamenta Informaticae 143(1-2) (2016) 1–18.

[13] J. Bazan, S. Bazan-Socha, S. Buregwa-Czuma, P. Pardel, A. Skowron, B. Sokolowska. Classifiers based on data sets and domain knowledge: A rough set approach. In: A. Skowron, Z. Suraj (Eds.), Rough Sets and Intelligent Systems - Professor Zdzisław Pawlak in Memoriam, Springer-Verlag, Berlin Heidelberg, Intelligent Systems Reference Library, vol. 43. 2013, pp. 93–136.

[14] J. Bazan, S. Bazan-Socha, S. Buregwa-Czuma, P. Pardel, B. Sokolowska. Predicting the presence of serious coronary artery disease based on 24 hour holter ecg monitoring. In: M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), Proceedings of the FedCSIS'2012, Wrocław. 2012, pp. 279–286.

[15] J. Bazan, S. Bazan-Socha, S. Buregwa-Czuma, P. Pardel, B. Sokolowska. Prediction of coronary arteriosclerosis in stable coronary heart disease. In: S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, R. R. Yager (Eds.), Advances in Computational Intelligence, Springer, Communications in Computer and Information Science, vol. 298. 2012, pp. 550–559.

[16] J. Bazan, S. Buregwa-Czuma, A. Jankowski. A domain knowledge as a tool for improving classifiers. Fundamenta Informaticae 127(1-4) (2013) 495–511.

[17] J. G. Bazan. A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery 1: Methodology and Applications, Physica-Verlag, Heidelberg, Germany, Studies in Fuzziness and Soft Computing, vol. 18. 1998, pp. 321–365.

[18] J. G. Bazan. Methods of approximate reasoning for synthesis of decision algorithms. Ph.D. thesis, Warsaw University, Warsaw, Poland, 1999. In Polish.

[19] J. G. Bazan. Behavioral pattern identification through rough set modeling. Fundamenta Informaticae 72(1-3) (2006) 37–50.

[20] J. G. Bazan. Hierarchical classifiers for complex spatio-temporal concepts. Transactions on Rough Sets IX, LNCS 5390 (2008) 474–750.

[21] J. G. Bazan, P. Kruczek, S. Bazan-Socha, A. Skowron, J. Pietrzyk. Risk pattern identification in the treatment of infants with respiratory failure through rough set modeling. In: Proceedings of the Eleventh Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'06), July 2-7, Paris, France. 2006, pp. 2650–2657.

[22] J. G. Bazan, P. Kruczek, S. Bazan-Socha, A. Skowron, J. Pietrzyk. Rough set approach to behavioral pattern identification. Fundamenta Informaticae 75(1-4) (2007) 27–47.

[23] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, J. Wróblewski. Rough set algorithms in classification problems. In: L. Polkowski, T. Y. Lin, S. Tsumoto (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Springer-Verlag/Physica-Verlag, Heidelberg, Germany, Studies in Fuzziness and Soft Computing, vol. 56. 2000, pp. 49–88.

[24] J. G. Bazan, A. Osmólski, A. Skowron, D. Ślęzak, M. Szczuka, J. Wróblewski. Rough set approach to the survival analysis. In: J. J. Alpigini, J. F. Peters, A. Skowron, N. Zhong (Eds.), Third International Conference on Rough Sets and Current Trends in Computing (RSCTC'02), October 14–16, Malvern, PA, USA. Springer-Verlag, London, UK, 2002, Lecture Notes in Artificial Intelligence, vol. 2475, pp. 522–529.

[25] J. G. Bazan, M. Szczuka. The Rough Set Exploration System. Transactions on Rough Sets 3400(3) (2005) 37–56.

[26] J. T. Bigger, J. L. Fleiss, R. C. Steinman, L. M. Rolnitzky, R. E. Kleiger, J. N. Rottman. Frequency domain measures of heart period variability and mortality after myocardial infarction. Circulation 85(1) (1992) 164–171.

[27] R. Brachman, T. Anand. The process of knowledge discovery in databases: A human-centered approach. Advances in Knowledge Discovery & Data Mining, AAAI Press & The MIT Press (1996) 37–57.

[28] R. Brachman, H. Levesque. Readings in Knowledge Representation. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1985.

[29] J. Brdulak. Zarządzanie wiedzą a proces innowacji produktu. Budowanie przewagi konkurencyjnej firmy. Szkoła Główna Handlowa, Warszawa, 2005.

[30] I. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. Classification and Regression Trees. Wadsworth, Belmond, CA, USA, 1984.

[31] BTL Polska Sp. z o.o. Strona internetowa: `http://www.btlnet.pl`.

[32] H. Bucher, P. Hengstler, C. Schindler. Percutaneous transluminal coronary angioplasty versus medical treatment for non-acute coronary heart disease: meta-analysis of randomised controlled trials. BMJ 321 (2000) 73–77.

[33] S. Buregwa-Czuma, J. Bazan, L. Zareba, S. Bazan-Socha, P. Pardel, B. Rewerska, L. Dydo. The method for describing changes in the perception of stenosis in blood vessels caused by an additional drug. Fundamenta Informaticae 147(2-3) (2016) 193–207.

[34] S. Buregwa-Czuma, J. Bazan, L. Zareba, S. Bazan-Socha, P. Pardel, B. Sokolowska, L. Dydo. The method for describing changes in the perception of stenosis in blood vessels caused by an additional drug. In: Proceedings of the 24th International Workshop on CS&P. Rzeszow, Poland. 2015, pp. 115–125.

[35] L. Cao, P. S. Yu, C. Zhang, H. Zhang. Data Mining for Business Applications. Springer US, Boston, MA, 2009.

[36] L. Cao, C. Zhang. Domain-driven data mining: a practical methodology. Int. J. of Data Warehousing and Mining 2(4) (2006) 49–65.

[37] C. Chien, L. Chen. Data mining to improve personnel selection and enhance human capital: A case study in hightechnology industry. Expert Systems with Applications 34 (2008) 280–290.

[38] Z. Chlewiński. Umysł. Dynamiczna organizacja pojęć. Wydawnictwo Naukowe PWN, Warszawa, 2000.

[39] P. Cichosz. Systemy uczące się. Wydawnictwa Naukowo-Techniczne, Warszawa, 2000.

[40] K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan. Data Mining. A Knowledge Discovery Approach. Springer US, Boston, MA, 2007.

[41] F. Crea, P. Camici, R. De Caterina, G. Lanza. Przewlekła choroba niedokrwienna serca. In: A. Camm, T. Luescher, P. Serruys (Eds.), Choroby serca i naczyń. Podręcznik Europejskiego Towarzystwa Kardiologicznego, Termedia Wydawnictwo Medyczne, Poznań, vol. I. 2006, pp. 409–444.

[42] R. Dalf. Organization Theory and Design. West Publishing Company, 2001.

[43] M. Dash, H. Liu. Feature selection for classification. Intelligent Data Analysis 1 (1997) 131–156.

[44] R. Davies, A. Goldberg, S. Forman. Asymptomatic cardiac ischaemia pilot (acip) study 2 year follow-up: outcomes of patients randomized to initial strategies of medical therapy versus revascularization. Circulation 95 (1997) 2037–2043.

[45] A. Desai. Adaptive complex enterprises. Communications ACM 5(48) (2005) 32–35.

[46] M. M. Deza, E. Deza. Encyclopedia of Distances. Springer Berlin Heidelberg, 2009.

[47] P. Domingos. The role of occam's razor in knowledge discovery. Data Mining and Knowledge Discovery, Kluwer Academic Publishers 3 (1999) 409–425.

[48] P. Domingos. Toward knowledge-rich data mining. Data Mining and Knowledge Discovery 1(15) (2007) 21–28.

[49] R. Dybowski, K. Laskey, J. Myers, S. Parsons. Introduction to the special issue on the fusion of domain knowledge with data for decision support. Journal of Machine Learning Research 4 (2003) 293–294.

[50] L. Dydo, J. G. Bazan, S. Buregwa-Czuma, W. Rzasa, A. Skowron. Verifying cuts as a tool for improving a classifier based on a decision tree. In: M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), Proceedings of the FedCSIS. 2016, ACSIS, vol. 8, pp. 17–20.

[51] T. Fawcett. An introduction to ROC analysis. Pattern Recognition Letters 27(8) (2006) 861–874.

[52] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery in databases. AI Magazine 17(3) (1996) 37–54.

[53] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In: Proc. of KDD-96. 1996, pp. 82–88.

[54] E. Folland, P. Hartigan, A. Parisi. Percutaneous transluminal coronary angioplasty versus medical therapy for stable angina pectoris: outcomes for patients with double-vessel versus single-vessel coronary artery disease in a veterans affairs cooperative randomized trial. veterans affairs acme investigators. J. Am. Coll. Cardiol. 29 (1997) 1505–1511.

[55] E. S. Ford, W. H. Giles, A. H. Mokdad. The distribution of 10-year risk for coronary heart disease among u.s. adults. Journal of the American College of Cardiology 43(10) (2004) 1791–1796.

[56] G. G. Gensini, C. Buonanno. Coronary arteriography: A study of 100 cases with angiographically proved coronary artery disease. Dis Chest 54 (1968) 90–93.

[57] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Smith, P. Sorlie, N. J. Stone, P. W. F. Wilson. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. Circulation 129(25 suppl 2) (2014) S49–S73.

[58] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286(5439) (1999) 531–537.

[59] G. Gordon, R. Jensen, L. Hsiao, S. Gullans, J. Blumenstock, S. Ramaswamy, W. Richards, D. Sugarbaker, R. Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research 62(17) (2002) 4963–4967.

[60] A. P. Gorgels, M. A. Vos, R. Mulleneers, C. de Zwaan, F. W. Bár, H. J. Wellens. Value of the electrocardiogram in diagnosing the number of severely narrowed coronary arteries in rest angina pectoris. Am J Cardiol. 72(14) (1993) 999–1003.

[61] T. R. Gruber. A translation approach to portable ontologies. Knowledge Acquisition 5(2) (1993) 199–220.

[62] N. Guarino. Formal ontology and information systems. In: Proceedings of the First International Conference on Formal Ontology in Information Systems (FOIS'98), June 6-8, Trento, Italy. IOS Press, 1998, pp. 3–15.

[63] J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd ed., 2011.

[64] D. Hand, H. Mannila, P. Smyth. Principles of Data Mining. MIT Press, Cambidge, MA, 2001.

[65] T. J. Hastie, R. J. Tibshirani, J. H. Friedman. The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics, Springer, New York, 2009.

[66] G. Haycock, G. Schwartz, D. Wisotsky. Geometric method for measuring body surface area: a height-weight formula validated in infants, children, and adults. J Pediatr. 93(1) (1978) 62–66.

[67] E. Heit. Background knowledge and models of categorization. In: U. Hahn, M. Ramscar (Eds.), Similarity and Categorization, Oxford University Press, USA, New York. 2000, pp. 155–178.

[68] R. Henderson, S. Pocock, T. Clayton, R. Knight, K. Fox, D. Julian, D. Chamberlain. Seven-year outcome in the rita-2 trial: coronary angioplasty versus medical therapy. J Am Coll Cardiol. 42(7) (2003) 1161–1170.

[69] H. Hirsh, M. Noordewier. Using background knowledge to improve inductive learning: a case study in molecular biology. IEEE Expert 10 (1994) 3–6.

[70] C.-C. Hsu, C.-L. Chen, Y.-W. Su. Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences 177 (2007) 4474–4492.

[71] Z. Huang, H. Chen, C. Hsu, W. Chen, S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems 7(4) (2004) 543–558.

[72] J. Jagielski. Inżynieria wiedzy w systemach ekspertowych. Lubuskie Towarzystwo Naukowe, Zielona Góra, 2001.

[73] M. Jarrar. Towards Methodological Principles for Ontology Engineering. Ph.D. thesis, Vrije Universiteit Brussel, 2005.

[74] P. Johnson. What kind of expert should a system be? Journal of Medicine and Philosophy 8(1) (1983) 77–97.

[75] D. Jones, T. Bench-Capon, P. Visser. Ontology-based support for human disease study. In: Proceedings of the IT&KNOWS Conference, XV IFIP World Computer Congress, August. 1998.

[76] W. Kannel, N. Feinleib. Natural history of angina in the framingham study. Am. J. Cardiol. 29 (1972) 154–165.

[77] A. Kaplan, G. Murphy. Category learning with minimal prior knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition 26(4) (2000) 829–846.

[78] R. Keefe. Theories of Vagueness. Cambridge University Press, New York, 2000.

[79] R. Keefe, P. Smith. Vagueness: A Reader. MIT Press, Massachusetts, MA, 1997.

[80] Kent Ridge Biomedical Dataset Repository. Strona internetowa projektu: `http://datam.i2r.a-star.edu.sg/datasets/krbd/`.

[81] Z. J. Klonowski. Systemy Informatyczne Zarządzania Przedsiębiorstwem. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2004.

[82] I. Kopanas, N. M. Avouris, S. Daskalaki. The role of domain knowledge in a large scale data mining project. In: I. Vlahavas, C. Spyropoulos (Eds.), SETN 2002, Springer-Verlag, Berlin Heidelberg, LNAI, vol. 2308. 2002, pp. 288–299.

[83] H. P. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, A. Zimek. Future trends in data mining. Data Mining and Knowledge Discovery 1(15) (2007) 87–97.

[84] M. Lam. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. Decision Support Systems 37(4) (2004) 567–581.

[85] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the Third SIAM International Conference on Data Mining, May 1-3, San Francisco, CA, USA. SIAM, 2003.

[86] C. Leacock, M. Chodorow. Combining local context and wordnet similarity for word sense identification. In: C. Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, Cambridge, MA. 1998, pp. 265–283.

[87] A. Liliana. Material and formal ontology. In: P. Roberto, S. Peter (Eds.), Formal ontology, Advanced in Soft Computing, Kluwer, Dordrecht, The Netherlands. 1996, pp. 199–232.

[88] D. Lin. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, ICML 1998. Morgan Kaufmann, Madison, WI, 1998, pp. 296–304.

[89] P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind. Geographic Information Systems and Science. Wiley, 2010.

[90] A. Maedche. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, Boston, MA, 2002.

[91] S. Mahmoodzadeh, M. Moazenzadeh, H. Rashidinejad, M. Sheikhvatan. Diagnostic performance of electrocardiography in the assessment of significant coronary artery disease and its anatomical size in comparison with coronary angiography. J. Res. Med. Sci. 16(6) (2011) 750–755.

[92] O. Maimon, L. Rokach. Data Mining and Knowledge Discovery Handbook. Springer Publishing Company, Incorporated, 2nd ed., 2010.

[93] C. J. Matheus, P. K. Chan, G. Piatetsky-Shapiro. Systems for knowledge discovery in databases. IEEE Transactions On Knowledge And Data Engineering 5 (1993) 903–913.

[94] R. Michalski, et al. (Eds.). Machine Learning, vol. I-IV. Morgan Kaufmann, Los Altos, 1983, 1986, 1990, 1994.

[95] D. Michie, D. J. Spiegelhalter, C. C. Taylor. Machine learning, neural and statistical classification. Ellis Horwood Limited, England, 1994.

[96] D. Morrow, B. Gersh, E. Braunwald. Przewlekła choroba wieńcowa. In: E. Braunwald, D. Zipes, L. P. (Eds.), Choroby serca, Urban and Partner, Wrocław. 2007, pp. 1251–1324.

[97] M. Motwani, J. Rana, R. Jain. Use of domain knowledge for fast mining of association rules. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists (IMECS). Hong Kong, 2009, vol. 1, pp. 18–20.

[98] J. Mulawka. Systemy ekspertowe. Wydawnictwa Naukowo-Techniczne, 1996.

[99] K. Napierala, J. Stefanowski. In: RSCTC. Springer-Verlag, 2010, Lecture Notes in Artificial Intelligence, vol. 6086, pp. 138–147.

[100] H. Nguyen, G. Rogers. Statistical inference, Fundamentals of Mathematical statistics, vol. II. Springer Verlag, New York, 1989.

[101] H. S. Nguyen. Approximate boolean reasoning: Foundations and applications in data mining. LNCS Transactions on Rough Sets V 4100 (2006) 334–506.

[102] N. F. Noy, D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical report, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[103] M. M. O. Owrang. Using domain knowledge to optimize the knowledge discovery process in databases. International Journal Of Intelligent Systems 15 (2000) 45–60.

[104] Z. Pawlak. Information systems - theoretical foundations. Information Systems 6 (1981) 205–218.

[105] Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data, D: System Theory, Knowledge Engineering and Problem Solving, vol. 9. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[106] Z. Pawlak, A. Skowron. A rough set approach for decision rules generation. In: Proceedings of Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93). Morgan Kaufmann, Chambéry, France, 1993, pp. 114–119.

[107] Z. Pawlak, A. Skowron. Rudiments of rough sets. Information Sciences 177 (2007) 3–27.

[108] M. Pazzani, D. Kibler. The utility of knowledge in inductive learning. Machine Learning 9(1) (1992) 57–94.

[109] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, C. Brunk. Reducing misclassification costs. In: Proceedings of the 11th Conference on Machine Learning. Rutgers Univ., New Brunswick, NJ, 1994, pp. 217–225.

[110] T. Pedersen, S. V. Pakhomov, S. Patwardhan, C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40 (2007) 288–299.

[111] J. F. Peters. Time and clock information systems: Concepts and rough fuzzy petri net models. In: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery 2. Applications, Case Studies and Software Systems, Studies in Fuzziness and Soft Computing, Springer-Verlag, Berlin. 1998, pp. 385–417.

[112] E. Petricoin, A. Ardekani, B. Hitt, P. Levine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, L. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 359 (2002) 572–577.

[113] G. Piatetsky-Shapiro, W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, Palo Alto, CA, USA, 1991.

[114] T. Poggio, S. Smale. The mathematics of learning: Dealing with data. Notices of the American Mathematical Society (AMS) 5(50) (2003) 537–544.

[115] C. Pohle. Integrating and updating domain knowledge with data mining. In: M. Scholl, T. Grust (Eds.), Proceedings of the VLDB 2003 PhD Workshop. Berlin, Germany, 2003.

[116] PostgreSQL Global Development Group. Strona internetowa: `https://www.postgresql.org`.

[117] S. G. Priori, E. Aliot, C. Blomstrom-Lundqvist, L. Bossaert, G. Breithardt, P. Brugada, A. J. Camm, R. Cappato, S. M. Cobbe, C. D. Mario, B. J. Maron, W. J. McKenna, A. K. Pedersen, U. Ravens, P. J. Schwartz, M. Trusz-Gluza, P. Vardas, H. J. J. Wellens, D. P. Zipes. Task force on sudden cardiac death of the european society of cardiology. Technical report, European Heart Journal, 2001.

[118] F. Provost, R. Kohavi. On applied research in machine learning. Machine Learning 30 (1998) 127– 132.

[119] J. Quinlan. Induction of decision trees. Machine Learning 1(1) (1986) 81–106.

[120] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California, 1993.

[121] R. Rada, H. Mili, E. Bicknell, M. Blettner. Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. 19(1) (1989) 17–30.

[122] F. Ravandi-Kashani, T. Hayes. Male breast cancer: a review of the literature. Eur J Cancer 34(1) (1998) 1341–1347.

[123] S. Read. Thinking about Logic: An Introduction to the Philosophy of Logic. Oxford University Press, New York, 1994.

[124] S. Redouane, R. S. Ahmed, B. Othmane. The integration of user knowledge to learn a specialized decision tree from a real-life data: an empirical and computational study. International Journal of Computer Science Issues 9(1) (2012) 310–317.

[125] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95. Morgan Kaufmann Publishers Inc., San-Francisco, CA, USA, 1995, vol. 1, pp. 448–453.

[126] B. R.M., B. M. J. R.L. Myocardial infarction with normal coronary arteries: A role for mri? Clin. Chem. 5 (2007) 995–996.

[127] S. P. Robbins, T. A. Judge. Organizational Behavior, 15th Edition. Prentice Hall, Prentice Hall, Person, Boston, 2012.

[128] RSES. Strona internetowa: `logic.mimuw.edu.pl/~rses`.

[129] Y. Ryu, W. Yue. Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches. IEEE Transactions on Systems, Man and Cybernetics Part A 35(5) (2005) 727–737.

[130] M. Sawicka-Parobczyk, K. Bieganowska. Odstęp QT/QTC w elektrokardiograficznym zapisie - ważny parametr, trudna ocena. Forum Medycyny Rodzinnej 4(1) (2010) 17–25.

[131] B. Scholkopf, A. J. Smola. Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, 2002.

[132] C. Shannon. A mathematical theory of communication. Bell System Technical Journal 27 (1948) 379–423.

[133] S. Sharma, K. Osei-Bryson. Data Mining for Business Applications, Springer US, Boston, MA, chap. Role of Human Intelligence in Domain Driven Data Mining. 2009, pp. 53–61.

[134] E. Shortliffe, B. Buchanan. A model of inexact reasoning in medicine. Mathematical Biosciences 23 (1975) 351–379.

[135] Y. Sinai. Probability theory. An introductory course. Springer Verlag, Berlin, 1992.

[136] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, W. Sellers. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1 (2002) 203–209.

[137] V. Singh, S. Nagpal. Integrating user's domain knowledge with association rule mining. International Journal of Computer Science Issues 7(5) (2010) 26–30.

[138] A. P. Sinha, H. Zhao. Incorporating domain knowledge into data mining classifiers: An application in indirect lending. Decision Support Systems 46 (2008) 287–299.

[139] A. Skowron, P. Synak. Complex patterns. Fundamenta Informaticae 60(1-4) (2004) 351–366.

[140] B. Smith. Ontology and information systems. Technical report, The Buffalo Ontology Site, `http://ontology.buffalo.edu/ontology(PIC).pdf`, 2001.

[141] P. Sneath, R. Sokal. Numerical Taxonomy. W.H. Freeman&Co., San Francisco, 1973.

[142] SOFA, Simple Ontology Framework API. Strona internetowa: `http://sofa.projects.semwebcentral.org`.

[143] B. Sokołowska. Wpływ hamowania biosyntezy leukotrienów na czynność elektryczną serca w stabilnej chorobie wieńcowej. Praca doktorska, Uniwersytet Jagielloński Collegium Medicum Wydział Lekarski, Kraków, 2010.

[144] S. Staab, R. E. Studer. Handbook on Ontologies. International Handbooks on Information Systems, Springer-Verlag Berlin Heidelberg, 2009.

[145] C. Stanfill, D. Waltz. Toward memory-based reasoning. Communications of the ACM 29 (1986) 1213–1228.

[146] STATISTICA (data analysis software system), version 10, StatSoft, Inc. Strona internetowa: `www.statsoft.com`.

[147] B. Stefanowicz. Informacja, wiedza, mądrość, Biblioteka Wiadomości Statystycznych, vol. 66. Zakład Wydawnictw Statystycznych, Warszawa, 2013.

[148] S. M. Straus, J. A. Kors, M. L. De Bruin, C. S. van der Hooft, A. Hofman, J. Heeringa, J. W. Deckers, J. H. Kingma, M. C. Sturkenboom, B. H. Stricker, J. C. Witteman. Prolonged QTc interval and risk of sudden cardiac death in a population of older adults. Journal of the American College of Cardiology 47(2) (2006) 362–367.

[149] J. A. Swets. Measuring the accuracy of diagnostic systems. Science 240 (1988) 1285–1293.

[150] P. Synak. Temporal Aspects of Data Analysis: A Rough Set Approach. Ph.D. thesis, The Institute of Computer Science of the Polish Academy of Sciences, Warsaw, Poland, 2003. In Polish, defended in 2004.

[151] A. Szczeklik, E. Nizankowska, L. Mastalerz, G. Bochenek. Myocardial ischemia possibly mediated by cysteinyl leukotrienes. J Allergy Clin Immunol 109 (2002) 572–573.

[152] M. A. H. Taieb, M. B. Aouicha, A. B. Hamadou. Ontology-based approach for measuring semantic similarity. Engineering Applications of Artificial Intelligence 36 (2014) 238–261.

[153] Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology. Heart Rate Variability. Standards of Measurement, Physiological Interpretation, and Clinical Use. Circulation 93(5) (1996) 1043–1065.

[154] The Elements of Statistical Learning Repository. Strona internetowa: `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/`.

[155] The Rough Sets Data Explorer (ROSE2). Strona internetowa projektu: `http://idss.cs.put.poznan.pl/site/rose.html`.

[156] The Weka 3, Data Mining Software in Java (WEKA). Strona internetowa: `http://www.cs.waikato.ac.nz/ml/weka/`.

[157] W. Traczyk. Structural representation of unstructured knowledge. Journal of Telecommunications and Information Technology 3 (2005) 81–86.

[158] UC Irvine Machine Learning Repository. Strona internetowa projektu: `http://archive.ics.uci.edu/ml/`.

[159] M. Uschold. Building ontologies: Towards a unified methodology. In: Proceedings 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems. Cambridge, UK, 1996.

[160] M. Uschold, M. Grüninger. Ontologies: principles, methods, and applications. Knowledge Engineering Review 11(2) (1996) 93–155.

[161] W3C. OWL Web Ontology Language, use cases and requirements, W3C recommendation. Technical report, The World Wide Web Consortium Technical Report, `http://www.w3.org/TR/2004/REC-webont-req-20040210/`, 2004.

[162] M. O. Ward, G. Grinstein, D. Keim. Interactive Data Visualization: Foundations, Techniques, and Applications. A K Peters/CRC Press, Natick, Massachusetts, 2015.

[163] M. Ware, E. Frank, G. Holmes, M. Hall, I. H. Witten. Interactive machine learning: letting users build classifiers. Int. J. Human-Computer Studies 55 (2001) 281–292.

[164] D. A. Waterman. A Guide to Expert Systems. Addison-Wesley, Reading, Mass., 1986.

[165] S. Weiss, S. Buckley, S. Kapoor, S. Damgaard. Knowledge-based data mining. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC., 2003, pp. 456–461.

[166] D. Wilson, T. Martinez. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6(1) (1997) 1–34.

[167] E. Wisniewski, D. Medin. On the interaction of theory and data in concept learning. Cognitive Science 18 (1994) 221–281.

[168] I. H. Witten, E. Frank, M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd ed., 2011.

[169] WordNet. Project web site: `http://wordnet.princeton.edu/`.

[170] Z. Wu, M. Palmer. Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL'94, Association for Computational Linguistics. Springer, Stroudsburg, PA, USA, 1994, vol. 4259, pp. 133–138.

[171] S.-C. Yoon, L. J. Henschen, E. K. Park, S. Makki. Using domain knowledge in knowledge discovery. In: Proc. ACM Conf. CIKM '99. Kansas City, MO, USA, 1999, pp. 243–250.

[172] S. Yusuf, D. Zucker, P. Peduzzi. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the coronary artery bypass graft surgery trialists collaboration. Lancet 344 (1994) 563–570.

[173] L. A. Zadeh. A new direction in AI: Toward a computational theory of perceptions. AI Magazine 22(1) (2004) 74–84.

[174] M. J. Zaki, M. Wagner, Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, New York, NY, USA, 2014.

[175] J. Zhang, D. Caragea, V. Honavar. Learning ontology-aware classifiers. In: A. Hoffmann, H. Motoda, T. Scheffer (Eds.), DS 2005, LNAI 3735, Springer-Verlag, Berlin Heidelberg. 2005, pp. 308–321.

[176] J. Zhang, A. Silvescu, V. Honavar. Ontology-driven induction of decision trees at multiple levels of abstraction. In: S. Koenig, R. Holte (Eds.), SARA 2002, LNAI 2371, Springer-Verlag, Berlin Heidelberg. 2002, pp. 316–323.

[177] H. Zhao, A. P. Sinha, W. Ge. Effects of feature construction on classification performance: An empirical study in bank failure prediction. Expert Systems with Applications 36 (2009) 2633–2644.

[178] D. Zhu, G. Premkumar, X. Zhang, C. Chu. Data mining for network intrusion detection: a comparison of alternative methods. Decision Sciences 32(4) (2001) 635–660.

[179] W. Ziarko. The discovery, analysis and representation of data dependencies in databases. In: G. Piatetsky-Shapiro, W. J. Frawley (Eds.), Knowledge Discovery in Databases, AAAI/MIT Press, Palo Alto, CA, USA. 1991, pp. 195–212.

Interdyscyplinarne Centrum Modelowania Komputerowego
Wydział Matematyczno-Przyrodniczy
Uniwerystet Rzeszowski

Skład przy użyciu systemu LaTeX.

BibTeX:

```
@thesis{key,
    author = "Sylwia Buregwa-Czuma",
    title = "{Metody stosowania wiedzy dziedzinowej do poprawiania jakości klasyfikatorów}",
    school = "Rzeszow University",
    address = "Rzeszow, Poland",
    year = "2017",
}
```