Rozdział 3

Wiedza i formy jej reprezentacji

Celem tego rozdziału jest szczegółowe określenie różnych form reprezentacji wiedzy, która stanowi podstawowy element każdego SE. Zaproponowany przez autora diagnostyczny SE bazuje na zintegrowaniu wiedzy o diagnozowanym obiekcie reprezentowanej w różnej formie. Zagadnienia zawarte w tym rozdziale obejmują również problematykę wydobywania wiedzy oraz eksploatacji hurtowni danych w diagnostyce.

3.1. Wstęp

Wiedza definiowana jest jako znajomość zjawisk zachodzących w środowisku, obiektach, zachowaniach ludzkich i zwierzęcych. Ponadto wiedza to także umiejętność opisu budowy, własności, parametrów obiektów z otoczenia człowieka i jego zakresu aktywności. Wiedza to także nabyte umiejętności zachowań, postępowania w aktywności codziennej oraz w sytuacjach wyjątkowych.

Ważnym zagadnieniem rozwiązywanym w czasie projektowania systemu ekspertowego jest reprezentacja wiedzy. Ze względu na charakter gromadzonej, przez człowieka i nie tylko (maszyny uczące się), wiedzy spotyka się różne formy reprezentacji wiedzy. Klasyczne opisy oparte na klasycznej logice formalnej (zdań) nie zawsze, a nawet coraz częściej, nie mogą być stosowane ze względu na niepewność i niejednoznaczność opisywanej wiedzy dotyczącej złożonych dynamicznych procesów.

3.2. Reprezentacja wiedzy

Reprezentacja wiedzy w danym systemie informacyjnym jest ważnym, a zarazem trudnym problemem, który nie został jeszcze w pełni rozwiązany. Wiedzę można reprezentować w formie *symbolicznej* lub *niesymbolicznej* i przedstawiać ją jako:

- *opisy* zawierają pierwotne cechy i pojęcia przedstawione w jakimś języku,
- *relacje* przedstawiają zależności i asocjacje pomiędzy faktami w bazie wiedzy.

Wyróżnia się dwa rodzaje *symbolicznej* reprezentacji wiedzy:

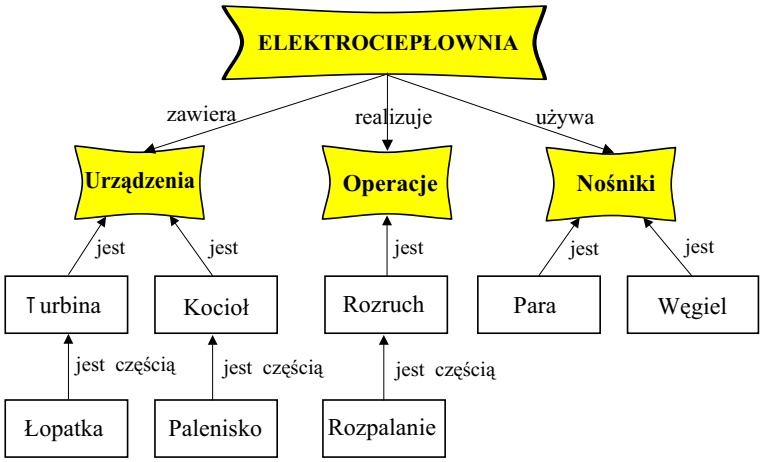
◇ *proceduralna* - polega na określeniu zbioru procedur, działanie których reprezentuje wiedzę o danej dziedzinie. Zaletą tej reprezentacji jest duża efektywność reprezentowania procesów, wymaga jednak znajomości praw fizyki opisanych za pomocą równań matematycznych lub dobrej znajomości funkcjonalnych powiązań między wybranymi sygnałami opisującymi stan procesu;

◇ *deklaratywna* - zawiera reguły empiryczne jako środek reprezentujący relacje pomiędzy cechami warunków działania i symptomami stanu a cechami stanu obiektu. Jest to mniej precyzyjna forma reprezentacji wiedzy o stanie obiektu [241].

Reprezentacje *niesymboliczne* bazują na obserwacji i doświadczeniach zebranych z otaczającego świata. Reprezentacje niesymboliczne realizuje się za pomocą technik opartych o elementy sztucznej inteligencji, do których między innymi można zaliczyć: sztuczne sieci neuronowe, algorytmy genetyczne lub logikę rozmytą.

Baza wiedzy to zbiór definicji, faktów, pojęć i relacji między nimi oraz reguł wnioskowania. Proces organizowania zebranej wiedzy wiąże się z wyborem odpowiedniej metody reprezentacji wiedzy oraz weryfikacji bazy wiedzy i mechanizmu wnioskowania. Do najczęściej stosowanych form reprezentowania wiedzy należy zaliczyć [8, 15]:

- **sieci semantyczne** - określają relacje pomiędzy elementami dziedziny (węzłami sieci). Realizują to poprzez definiowanie połączeń pomiędzy węzłami (łuki). Węzły reprezentują zwykle obiekty fizyczne lub koncepcyjne oraz deskryptory (cechy charakterystyczne obiektów). Łuki łączą obiekty i ich deskryptory. Główną zaletą tej formy reprezentacji wiedzy jest jej elastyczność, przez dziedziczenie cech nadrzędnej klasy obiektów, przez podrzędne elementy tej klasy (rys. 3.1).
- **trójki** - $<$ *obiekt, atrybut, wartość* $>$. Ta forma reprezentacji wiedzy jest szczególnym przypadkiem sieci semantycznej. Atrybuty są generalnymi cechami obiektów. W tej formie opracowano bazę wiedzy systemu ekspertowego



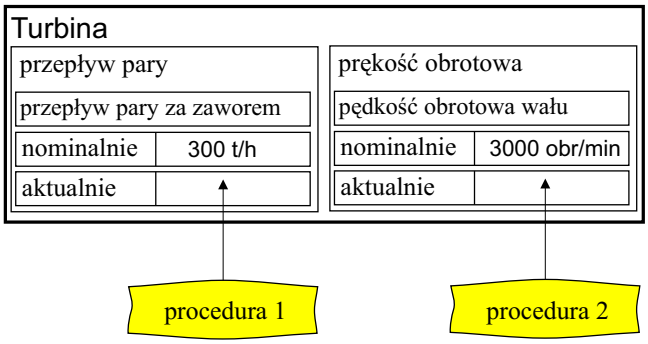
Rys. 3.1. Struktura przykładowej sieci semantycznej

stosowanego w medycynie MYCIN.
Forma opisu wiedzy z zastosowaniem trójek zawierających definicję poziomu ufności wiedzy może być przedstawiona jako:

$$<o,\ n(x),\ val(x),\ CF>.$$

gdzie: o oznacza obiekt, $n(x)$ - cecha danego obiektu, $val(x)$ - wartość cechy, CF - poziomem ufności.
Przykładowo dla zespołu elektrowni cieplnej można wyróżnić pewną grupę trójek reprezentujących aktualny stan badanego zespołu, takich jak:

$$<\ para,\ temperatura,\ 793K,\ 0.7\ >,$$
$$<\ turbina,\ prędkość\ obrotowa,\ 3000,\ 0.86\ >.$$

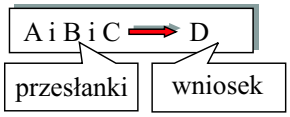


Rys. 3.2. Rama opisująca obiekt Turbina

- **reguły wnioskowania** - budowane są w formie powiązań między grupą przesłanek i wynikającą z nich grupą konkluzji. Przyjmują postać [163]: **jeżeli** $(w_1\ i\ w_2\ i\ \cdots\ i\ w_n)$ **to** $(d_1\ i\ d_2\ i\ \cdots\ i\ d_m)$, gdzie: $w_1, w_2, \cdots, w_n$ oznaczają warunki, które muszą być spełnione aby uaktywnić regułę, d_1, $d_2, \cdots, d_m$ - konkluzje określające działanie przy aktywnej regule.

- **ramy** - grupują dane i procedury w obiekty lub ramy. Oparte są na hierarchicznej strukturze dziedziczenia cech w dół. Stanowią one elastyczną wersję rekordu. Pola rekordu odpowiadają atrybutom obiektu. Łączone są zazwyczaj między sobą w strukturę hierarchiczną zgodnie z zasadą *"od ogółu do szczegółu"* (rys. 3.2) [158].

- **warunki Horna** - wyrażenia w formie predykatów. Struktura zbliżona do generalnej postaci reguł wnioskowania z dodatkowymi restrykcjami, wynikającymi z syntaktyki rachunku predykatów. Metoda ta była przyczynkiem do budowania niektórych języków programowania (Prolog).
  Warunki Horna budowane są w strukturze reguły o jednym wniosku (rys. 3.3) [163]. Baza wiedzy budowana w oparciu o warunki Horna wykazuje się:



Rys. 3.3. Przykład struktury warunku Horna

  - bardzo uproszczoną automatyzacją wnioskowania,
  - efektywnym mechanizmem wnioskowania,
  - prostymi, zrozumiałymi i przejrzystymi regułami.

  Do wad tej formy reprezentacji wiedzy należy duża liczba reguł stosowanych w złożonych bazach wiedzy.

3.3. Akwizycja wiedzy

Proces pozyskiwania wiedzy to pozyskanie zasobu wiedzy i doświadczenia, odpowiadających zakresowi zadań z danej dziedziny zastosowania, ze zidentyfikowanych źródeł wiedzy. Drugim etapem działania jest zapisanie ich w bazie wiedzy w sposób umożliwiający skuteczne wspomaganie działania człowieka podczas rozwiązywania problemów z tej dziedziny [153].
Istotnym elementem pozyskiwania wiedzy jest źródło tej wiedzy. Najważniejszymi źródłami wiedzy dla systemów ekspertowych są:

- specjaliści (eksperci), od których wiedzę uzyskuje się w sposób bezpośredni poprzez udział ich w procesie pozyskiwania lub w sposób pośredni, korzystając z literatury fachowej analizowanej przez osoby trzecie lub specjalistyczne oprogramowanie [242],
- bazy danych zawierające wyniki obserwacji diagnozowanego obiektu (grupy obiektów) lub wyniki obliczeń symulacyjnych prowadzonych z zastosowaniem odpowiedniego typu modelu.

Pozyskiwanie wiedzy od specjalistów historycznie jest metodą najwcześniej stosowaną. Ekspert lub grupa ekspertów przekazywała swoje doświadczenia, które następnie inżynier wiedzy, po odpowiednim przetworzeniu, wprowadzał do bazy wiedzy. Niekiedy sam ekspert pełnił funkcję inżyniera wiedzy. Obecnie również ta metoda pozyskiwania wiedzy jest zalecana szczególnie w procesie przygotowywania wstępnej wersji systemu ekspertowego. Jednakże wskazane jest również ograniczenie funkcji inżyniera wiedzy w procesie pozyskiwania wiedzy. Zaleca się stosowanie zaawansowanych programów komputerowych, które realizują funkcje odpytywania eksperta i uzupełniania bazy wiedzy w procesie samo-uczenia się systemu ekspertowego. Wyeliminowanie inżyniera wiedzy z procesu pozyskiwania wiedzy ma na celu ograniczenie funkcji osób trzecich nie zorientowanych w zagadnieniach diagnozowanego procesu. Ich wpływ na zawartość bazy wiedzy jest niekiedy bardzo duży, to oni interpretują informacje uzyskane podczas przeprowadzonych wywiadów z ekspertem [188].

Zadanie pozyskiwania wiedzy od eksperta jest czasami problemem z zakresu socjologii i psychologii. Niekiedy eksperci nie chcą uczestniczyć w czynnej formie pozyskiwania wiedzy (wywiady). W takich sytuacjach można zastosować bierny udział specjalisty. Polega on na analizie zarejestrowanych wypowiedzi specjalisty podczas rozwiązywania problemów diagnostycznych lub na obserwacji specjalisty podczas rozwiązywania przez niego zadań diagnostycznych. Istotnym elementem jest prowadzenie obserwacji w sposób ostrożny, aby to w jak najmniejszym stopniu nie wpływało na specjalistę.

Kolejnym problemem w procesie pozyskiwania wiedzy jest zadanie łączenia wiedzy pochodzącej od wielu ekspertów. Często uzyskuje się wtedy rozbieżne opinie w zakresie określania wartości stwierdzeń stosowanych do budowy elementarnych warunków przesłanek lub konkluzji reguł oraz określania *"stopnia pewności"* reguł. W tej sytuacji rozwiązanie tego problemu widzi się w procesie agregacji opinii pochodzących od wielu ekspertów. Zadanie to można rozwiązać, korzystając z doświadczenia inżyniera wiedzy lub za pomocą specjalistycznego oprogramowania realizującego proces specyficznego uczenia się systemu doradczego. W literaturze [92] można znaleźć informacje o trudnościach, jakie pojawiają się przy pozyskiwaniu wiedzy od wielu ekspertów. Biorąc to pod uwagę, należy stwierdzić, że często źródłem wiedzy dla danej bazy jest tylko jeden ekspert.

Realizacja zadania pozyskiwania wiedzy od ekspertów wiąże się z problemem obiektywności dialogu z ekspertem oraz wykonania etapu wydobywania wiedzy z danych zgromadzonych w czasie dialogu [92].

Pozyskiwanie wiedzy to przetworzenie danych uzyskanych różnymi metodami (np. w dialogu z ekspertem) w wymaganą (lub możliwą do uzyskania) wiedzę. Wiedza zgromadzona w procesie pozyskiwania wiedzy nie zawsze jest zgodna z wiedzą eksperta. Wynika to często ze złego zrozumienia przez eksperta zagadnienia lub nieprawidłowego przetworzenia informacji zgromadzonej w czasie wywiadu przez inżyniera wiedzy. Wydobywanie pozyskanej wiedzy wymaga konfrontacji ustaleń inżyniera wiedzy z ekspertem (weryfikacji bazy wiedzy opracowanej przez inżyniera wiedzy). Konfrontacja ta ma na celu upewnić się co do poprawności interpretacji wiedzy eksperta.

W procesie pozyskiwania wiedzy od ekspertów ważna jest zgodność ocen atrybutów przygotowanej bazy wiedzy przez grupę ekspertów. Ze statystycznego punktu widzenia ważny jest stopień korelacji między n - ocenami ekspertów dla m - atrybutów. Miarą tej współzależności jest współczynnik W-Kendalla [92]. Współczynnik ten przyjmuje wartość $W = 0$ przy braku zgodności i $W = 1$ przy pełnej zgodności. Duża wartość W oznacza dobrą zgodność ocen ekspertów, ale z tego nie można wnioskować o dużej poprawności przyjętego rozwiązania. Poprawność zależy od wiedzy ekspertów i uwarunkowań pracy diagnozowanego obiektu. Proces pozyskiwania wiedzy od ekspertów ma w sobie pewne cechy losowości. W związku z tym sama znajomość wartości współczynnika W nie jest wystarczającym kryterium oceny jakości wiedzy. Do oceny wartości statystycznej istotności na danym poziomie α należy skorzystać z testu χ^2 [92].

Druga grupa metod pozyskiwania wiedzy realizowana jest przy wykorzystaniu baz danych tworzonych w czasie eksploatacji diagnozowanego procesu lub grupy podobnych procesów. Proces pozyskiwania wiedzy z baz danych można realizować dwoma metodami:

- indukcyjną (maszynową) realizowaną na podstawie przykładów sklasyfikowanych przez nauczyciela,
- odkrywania zależności jakościowych i ilościowych (funkcyjnych) w bazach danych.

Metoda indukcyjna korzysta z tzw. modelu atrybutowego, w którym dane uczące zapisywane są w jednym zbiorze jako wartości cech opisujących właściwości obiektu znajdującego się w zadanym stanie oraz wartości cechy decyzyjnej opisującej stan obiektu. W ramach badań nad tym zagadnieniem można wyróżnić kilka metod, do których przede wszystkim należy zaliczyć:

- metodę realizowaną za pomocą pokryć według algorytmu AQ [151],

- generowanie drzew decyzyjnych [206],
- pozyskiwanie reguł oraz klasyfikatorów przybliżonych [251],
- metody oparte o sieci neuronowe [5, 107],
- metody oparte o logikę rozmytą [185].

Pozyskiwanie wiedzy metodami odkryć w bazach opracowano w ostatnim okresie [154]. Cechą charakterystyczną tych metod jest pozyskiwanie nowej wiedzy, podczas gdy w metodzie indukcyjnej realizuje się zadanie pozyskiwania wiedzy już wcześniej odkrytej, a celem procesu maszynowego uczenia jest odpowiednia reprezentacja jej w bazie. Źródłem użytecznej wiedzy o relacjach diagnostycznych mogą być diagnostyczne bazy danych. Zawierają one wartości cech, które opisują wejścia i wyjścia obserwowanych obiektów. Odkrywanie wiedzy opiera się na poszukiwaniu regularności występujących w stosowanym zbiorze danych.

3.4. Przetwarzanie baz danych jako forma wydobywania wiedzy

Zebrane w bazach danych zapisy zawierające informacje z różnych dziedzin działalności człowieka mogą być wykorzystane do wspomagania podejmowania decyzji o dalszym kształtowaniu profilu prowadzonej działalności. Komputerowe systemy wspomagania decyzji (ang. *Decision Support Systems*) bazują na zgromadzonej wiedzy ekspertów, częściowo pochodzącej z analizy zawartości baz danych.

Pierwsze procedury przetwarzania baz danych realizowano w oparciu o opracowany w latach siedemdziesiątych język strukturalny SQL (ang. *Structured Query Language*). Pozwalał on na budowę złożonych i skomplikowanych zapytań co umożliwiło wydobywać wybrane charakterystyczne cechy przetwarzanych danych. Następny przełom nastąpił w latach osiemdziesiątych poprzez wprowadzenie technologii nazywanej *Hurtownią Danych* (ang. *Data Warehouses*) [237].

Hurtownie danych miały za zadanie łączyć i kolekcjonować informacje pochodzące z różnych autonomicznych systemów oraz zewnętrznych źródeł. Pozwoliły one nie tylko uporządkować rozrzucone i nadmiarowe dane w postaci tzw. kostek (ang. *Cubes*) ale również dokonać wstępnej filtracji i przetwarzania potrzebnej informacji na podstawie projektu modelu danych.

Połączenie Hurtowni Danych i systemów wspomagania decyzji pozwoliło na stworzenie środowiska typu OLAP (ang. *OnLine Analytical Processing*). Umożliwia ono wielowymiarową obserwację agregowanych wartości wybranych atrybutów jednej lub wielu połączonych relacji. Metodologia tego środowiska zakłada, że użytkownik definiuje pewną hipotezę, której poprawność weryfikuje się za pomocą narzędzi OLAP (np. Oracle Express Server). Przedstawiona metoda ma ograniczenia

stosowalności i związaną z nimi skuteczność akwizycji wiedzy. Do najważniejszych należy zaliczyć konieczność przygotowania przez użytkownika hipotez do weryfikacji. Jakość wydobywanej z bazy danych wiedzy jest ograniczona kreatywnością i wyobraźnią eksperta definiującego hipotezy. Ponadto istnieje ryzyko akceptacji fałszywych hipotez [48].

Dynamiczny rozwój sztucznej inteligencji (koniec lat dziewięćdziesiątych) wyodrębnił nową dziedzinę pozyskiwania wiedzy nazywaną eksploracją wiedzy, DM. Aktualnie dynamicznie rozwijana jest dziedzina automatycznego odkrywania wiedzy w bazach danych (ang. *Knowledge Discovery in Databases, KDD*) i jej technologia eksploracji danych DM. Analizując rozwój dzisiejszych technologii odkrywania wiedzy z baz danych trudno jest określić, jak będą one wyglądały w przyszłości. Ostatnią z opracowanych technologii, które doczekały się komercyjnego zastosowania, jest metoda ułatwiająca dostęp do odkrytych schematów nazywana *dostęp do wiedzy* (ang. *Knowledge Access, KA*). W tej metodzie zastosowano nowe wcielenie języka SQL nazywanego PQL (ang. *The Pattern Query Language*), który swobodnie operuje wśród zgromadzonej bazy wzorców [48].

3.4.1. Odkrywanie wiedzy w bazach danych

Definicja 3.1.
KDD jest nietrywialnym procesem identyfikowania ważnych, nowych, potencjalnie użytecznych i zrozumiałych wzorców w danych [253].

Proces odkrywania wiedzy to wysoce intensywne współdziałanie człowieka i maszyny. Człowiek określa, które regularności są interesujące, nowe i zrozumiałe. Komputer maksymalnie przetwarza lawinę informacji zawartej w bazie danych przedstawiając najbardziej wiarygodne hipotezy [248]. W chwili obecnej prace koncentrują się na pozyskiwaniu wiedzy z danych, co jest jednym z elementów procesu KDD. Jednakże w coraz większym stopniu prowadzi się prace nad etapami przygotowania danych, wizualizacji oraz integracji uzyskanej wiedzy z istniejącymi systemami użytkownika. Należy tutaj zaznaczyć, że często można spotkać się z łączeniem pojęć odkrywania wiedzy (KDD) i eksploracji danych (DM). Nic bardziej mylnego, eksploracja danych jest jednym z elementów odkrywania wiedzy.

Poszczególne operacje realizowane podczas odkrywania wiedzy [237] definiuje się w następujący sposób:

- Wstępna selekcja danych w celu takiego doboru informacji, która jest dobrym nośnikiem poszukiwanej wiedzy;
- Przygotowanie danych polegające na uzupełnieniu braków oraz standaryzacji różnych formatów przechowywanych danych. Na tym etapie wykonuje się również czyszczenie danych z szumu i wyrzutków (odszumianie);

- ◇ Transformacja polega na konwersji danych na formę akceptowalną przez zastosowany przez operatora system wydobywania wiedzy;
- ◇ Eksploracja polegająca na poszukiwaniu odkrytych schematów. Operację tę wykonuje się przy wykorzystaniu różnych metod z nią związanych;
- ◇ Prezentacja wiedzy zamyka proces odkrywania wiedzy. Uzyskana wiedza może podlegać wizualizacji lub dalszej analizie w dowolnym programie wspomagającym wnioskowanie wyników.

Część procesu transformacji oraz procedury eksploracji i prezentacji należą do dziedziny eksploracji danych, DM.

3.4.2. Cele i techniki eksploracji danych

Proces eksploracji danych obejmuje dwa cele:

- *przewidywanie* (ang. *prediction*) obejmujące zadanie wykorzystania określonych danych do prognozowania wartości innych interesujących zmiennych,
- *opisywanie* (ang. *description*) cechuje się poszukiwaniem możliwych do zinterpretowania przez człowieka wzorców opisujących dane.

Eksploracja danych może być realizowana za pomocą różnych technik. Wybór techniki do prowadzenia tego procesu zależy od typu danych przechowywanych w analizowanej bazie oraz od schematów zdefiniowanych przez użytkownika. Najczęściej używane techniki można przedstawić jako:

- ◇ ***Klasyfikacja*** realizująca zadanie wyznaczenia funkcji, która odwzorowuje punkt danych w jedną z wielu *wcześniej zdefiniowanych* klas. Klasyfikacja polega na przyporządkowaniu klasyfikowanego elementu do klasy, do której odległość jest minimalna i można ją realizować w formie [30]:
  - *binarnej*,
  - *wieloklasowej*.

  Z punktu widzenia formy realizacji zadań klasyfikacji klasyfikatory można podzielić na statystyczne oraz oparte na bezpośrednim przyjmowaniu funkcji określającej klasyfikator. Rozważając algorytmy klasyfikacji wyróżnia się [180]:
  - *klasyfikację rozmytą*, w której wyznacza się stopień przynależności klasyfikowanego elementu do każdej klasy,
  - *klasyfikację dokładną* zaliczaną do szczególnego przypadku klasyfikacji rozmytej. W tym algorytmie zamiast stopnia przynależności określa się oznaczenie klasy, do której dany element został zakwalifikowany.

- ◇ ***Regresja*** polegająca na wyznaczeniu parametrów funkcji (liniowej lub nieliniowej), która dla określonej danej wyznacza wartość typu rzeczywistego. Przykładem zastosowań regresji może być przewidywanie wielkości temperatury silnika spalinowego na podstawie kilku wyników pomiarów jego punktów pracy [214].

- ◇ ***Grupowanie pojęciowe*** stosowane do znajdowania skończonego zbioru klas obiektów (znanych również jako: klastry, zbiory lub segmenty) w bazie danych posiadających podobne cechy.
  Ze względu na istotę działania wśród algorytmów grupowania można wyróżnić [30, 48]:

  - algorytmy poszukiwania ogólnego ekstremum funkcji kryterialnej,
  - algorytmy iteracyjne,
  - algorytmy hierarchicznego łączenia,
  - algorytmy oparte na teorii grafów: algorytm najbliższego sąsiada, algorytm k-najbliższych sąsiadów, algorytm minimalnego drzewa, algorytm MMD (ang. *Mean Minimum Distance*),
  - algorytmy wykorzystujące zbiory rozmyte (np. FCM).

- ◇ ***Kojarzenie*** (asocjacja) to poszukiwanie elementów, które wiążą się z zadanym zdarzeniem lub innym elementem. Stosowane do realizacji tego celu algorytmy pozwalają wyznaczyć reguły wiążące te elementy.

Analizując przedstawione techniki eksploracji danych można zauważyć, że *klasyfikacja* i *regresja* są najczęściej stosowane do tworzenia prognoz, czyli do przewidywania zdarzeń, natomiast *grupowanie* i *kojarzenie* doskonale nadaje się do opisu procesów zdefiniowanych za pomocą eksplorowanej bazy danych.

Każdy z przedstawionych algorytmów eksploracji danych składa się z trzech podstawowych składników:

- ◇ reprezentacji modelu (ang. *model representation*), zawierającej możliwe do odkrycia wzorce;

- ◇ oceny modelu(ang. *model evaluation*), stosowanej do oceny jakości otrzymanego w procesie KDD modelu;

- ◇ poszukiwania (ang. *searching*), obejmującego zagadnienia: poszukiwania parametrów (ang. *parameters search*) i poszukiwania modeli
  (ang. *model search*).

Proces *poszukiwania parametrów* realizuje zadanie doboru do zdefiniowanego modelu wartości parametrów tego modelu, które optymalizują kryteria oceny modelu dla analizowanych danych. W przypadku prostego modelu zadanie to można

zrealizować wykorzystując algorytmy optymalizacji jedno lub wieloparametrycznej (gradientowej lub bezgradientowej). Dla bardziej ogólnych modeli (w przypadku problemów NP-zupełnych) stosuje się również metody przeszukiwania heurystycznego.

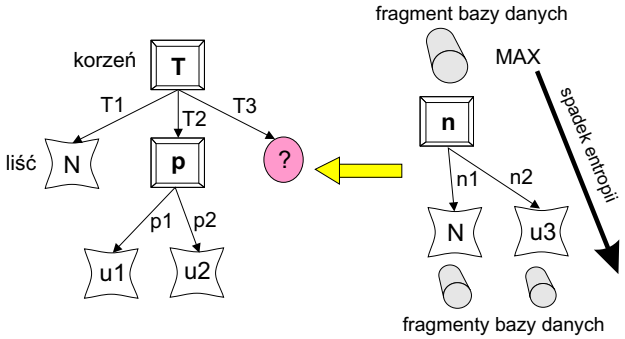
Poszukiwanie modeli polega na doborze parametrów dla grupy modeli lub doborze parametrów i struktury modelu. Implementacje metod poszukiwania modelu ewoluują w kierunku stosowania technik poszukiwania heurystycznego, ponieważ rozmiar problemu często wyklucza uzyskanie wyczerpującego poszukiwania metodami klasycznych technik optymalizacyjnych.

3.4.3. Typy algorytmów odkryć

Metody stosowane w różnych algorytmach eksploracji danych pozwalają następująco podzielić stosowane algorytmy [3, 39, 159]:

- drzewa i reguły decyzyjne [181],
- regresja liniowa, nieliniowa i klasyfikacja,
- wnioskowanie z przykładów,
- probabilistyczne modele graficznej zależności,
- relacyjne metody uczenia.

W **drzewach i regułach decyzyjnych** wykorzystuje się podziały przestrzeni względem jednej zmiennej. Zmienne stosowane w algorytmie dzieli się na informacyjne i decyzyjną (klasyfikacyjną). Zmienna decyzyjna przyjmuje formę dyskretną. Liście tworzonego drzewa przyjmują wartość zmiennej decyzyjnej. Ścieżka od korzenia do danego liścia reprezentuje warunki, których spełnienie implikuje wartość zmiennej decyzyjnej równej zawartości liścia [230].



Rys. 3.4. Proces uczenia drzewa decyzyjnego z danych

Proces uczenia się drzew decyzyjnych z danych polega na tworzeniu łańcuchów zmiennych informacyjnych dzielących bazę danych na coraz mniejsze części (rys.3.4). Proces zostaje zakończony, gdy otrzymana część bazy danych jest jednorodna ze względu na zmienną klasyfikacyjną lub brak możliwości podniesienia stopnia jednorodności, lub otrzymany fragment bazy jest mało liczny. Algorytmy tego typu stosuje się najczęściej wtedy, gdy istnieje deterministyczna lub prawie deterministyczna zależność między zmiennymi informacyjnymi a zmienną klasyfikacyjną [24].

Regresja liniowa, nieliniowa i klasyfikacja obejmują grupę zagadnień dotyczących wyznaczenia dla badanych danych przebiegu funkcji $y = f(x_1, x_2, \cdots, x_p)$. W praktyce ze względu na występujące w procesie pomiarowym zakłócenia losowe nie spotyka się czystej zależności funkcyjnej, lecz zależność stochastyczną $Y = F(X_1, X_2, \cdots, X_p)$, gdzie X_i i Y są zmiennymi losowymi.

Proces poszukiwania rzeczywistej zależności pomiędzy określonymi zmiennymi losowymi można realizować wykorzystując funkcję regresji w postaci:

$$E\left(Y\big|_{X_1=x_1, X_2=x_2,\ldots}\right) = \mu_y = f(x_1, x_2, \cdots, x_p). \tag{3.1}$$

W przypadku regresji liniowej (w celu doświadczalnego ustalenia przebiegu funkcji regresji) przyjmuje się wielomianowy model funkcji w postaci:

$$\widehat{Y}_g = g\left(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p\right), \tag{3.2}$$

gdzie: $g(X)$ jest funkcją wektora $\mathbf{X} = [X_1, X_2, \cdots, X_p]$, a parametry b_i zwane są współczynnikami regresji.
Przy wyznaczaniu współczynników regresji przyjmuje się kryterium minimalizacji kwadratów odchyleń wyników od prostej regresji opisane zależnością:

$$SSE = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(y_{(i)} - \widehat{y}_{(i)}\right)^2 = min, \tag{3.3}$$

gdzie: $e_{(i)}$ opisuje tzw. błąd obserwacji lub resztę i oznacza odchylenie i-tej obserwacji od prostej regresji.

Nieliniowa regresja to szereg technik predykcji zmiennej wyjściowej z liniowej lub nieliniowej kombinacji funkcji bazowych (np. sigmoid, wielomianów itp.). Proces poszukiwania funkcji regresji rozpoczyna się od próby opisania całości danych za pomocą jednej funkcji. Jeśli przeprowadzona próba zakończy się niepowodzeniem, dokonuje się podziału danych na mniejsze grupy i dokonuje ponownej próby znalezienia funkcji regresji dla poszczególnych grup danych. Powoduje to poszatkowanie przestrzeni obserwacji, którą w kolejnym etapie techniką redukcji *pruning* fragmentami scala się. Ideą metod redukcji jest możliwie uprościć kształt granic

podziału przestrzeni wejściowej [47].

Wnioskowanie z przykładów opiera się na reprezentatywnych przykładach pobranych z analizowanej bazy danych. Na bazie przykładów aproksymuje się lokalnie model. Aproksymacji dokonuje się na bazie klasyfikacji lub regresji w oparciu o najbliższe otoczenie, bądź tzw. metodą systemów wnioskowania z przypadków. Poprawność funkcjonowania tej metody wymaga zdefiniowania odpowiedniej metryki. W trakcie uczenia określa się metrykę odległości oraz liczbę uwzględnianych sąsiadów.

Probabilistyczne modele graficznej zależności są reprezentowane przez sieci bayesowskie definiowane jako zwarta reprezentacja łącznego prawdopodobieństwa. Sieć bayesowska stanowi acykliczny graf skierowany, w którym węzły reprezentują zmienne, a łuki - bezpośrednie zależności przyczynowe. Łączny rozkład prawdopodobieństwa jest modelowany przez iloczyn warunkowych rozkładów zmiennych względem ich rodziców.

Relacyjne metody uczenia definiują relacje z bazy danych. Do budowy relacji wykorzystuje się niektóre instancje relacji oraz przypadki negatywne relacji. Zadanie budowy relacji wykonuje się w oparciu o metody indukcyjne ILP (ang. *Inductive Logic Programming*) realizowane jako:

- *empiryczne*, realizowane w pełni automatycznie z wsadowym konstruowaniu pojedynczych pojęć od zera;
- *interakcyjne*, pracujące na zasadzie współdziałania systemu z użytkownikiem przygotowując jednocześnie wiele definicji nowych pojęć.

3.5. Hurtownie danych w diagnostyce

Ostatnie 10 lat stanowi okres burzliwego rozwoju prac nad zasadami zarządzania i przetwarzania dużych rozproszonych lub skupionych baz danych. Rozwój dotyczy tendencji do przechowywania istotnych dla danego zagadnienia danych a nie przechowywania dużych ilości encyklopedycznych danych bez ponownego ich wykorzystania (tylko archiwum). Nowe osiągnięcia w tej dziedzinie dotyczą takich obszarów działalności człowieka jak gospodarka oraz nauka i prowadzą do strukturalizacji informacji. Archiwizowane dane, związane z funkcjonowaniem pewnej organizacji w pewnym czasie, mogą być ważnym elementem dla planowania strategicznego lub diagnostyki.

Szeroko znaną techniką wydobywania wiedzy jest statystyka. Stanowi ona tradycyjne pole wnioskowania dostarczające modeli przy założeniu mniej lub bardziej szczegółowego rozkładu danych. Klasyczna teoria wnioskowania Bayes'a zapewnia dobrą skuteczność przy niezbyt dużych zbiorach danych. Przy dużych zasobach

zmiennych z zakresu takich aplikacji jak medyczne, marketingowe, teoria Bayes'a wykazuje niewystarczającą skuteczność.

Nowe metody opierają się w mniejszym stopniu na założeniach statystycznych, a w większym na - aktualnej dystrybucji danych. Ponadto polegają one mniej na prostych modelach matematycznych a bardziej na modelach opartych o techniki sztucznej inteligencji, które można uczyć skomplikowanych, nieliniowych zależności, korzystając z dużych zbiorów danych [50].

Przygotowując aplikacje wydobycia wiedzy, należy zwracać uwagę na następujące istotne zagadnienia:

- *Duża wymiarowość zbioru danych.* Problem obejmuje zadanie, w którym liczba przykładów jest mała w porównaniu z liczbą danych występujących w bazie. Problem ten pojawia się często w przypadku stosowania dyskretnych zmiennych przy małym kroku dyskretyzacji. Przy modelowaniu rozkładu powiązań wielu losowych zmiennych za pomocą złożonych instrukcji pojawia się znane zagadnienie *przekleństwa wymiarowości* i uzyskuje się modele z wykładniczym wzrostem liczby poszukiwanych parametrów. W literaturze spotyka się różne propozycje rozwiązywania tego typu problemów. Bengio i Bengio [16] proponują problem przetwarzania danych w dużej wymiarowości rozwiązać na bazie sztucznych sieci neuronowych (ANN), które będą reprezentować rozkład powiązań wielu zmiennych. Innym podejściem do tego problemu jest uczenie modeli na bazie maksimum entropii. Przykład zastosowania tej metody opisano w artykule Yan i Miller [244], gdzie autorzy, bazując na zasadzie aproksymacji maksimum entropii, jako alternatywy dla sieci Bayes'a zbudowali ogólny mechanizm wnioskowania statystycznego. W artykule Kewley'a [100] zaprezentowano zastosowanie metod krzyżowania i analizy wrażliwości sieci do trudnych problemów analiz chemicznych, gdzie spotyka się problem większej liczby zmiennych niż danych uczących.

- *Skalowalność algorytmów do dużych zbiorów danych.* Wiele metod proponowanych ostatnio wymaga weryfikacji według kryterium możliwości ich zastosowań z bardzo dużymi zbiorami danych. W przypadku zastosowania technik, bazujących na uczeniu w oparciu o przykłady, występuje problem przeuczenia sieci. Pojawia się pytanie *jak uczyć sieć przy dużej ilości danych przy założeniu zachowania rozsądnego czasu obliczeń.*

- *Wydobywanie użytecznej informacji i budowa użytecznych aplikacji.* Zagadnienie to obejmuje zadania zrozumienia danych przez wydobywanie reguł, klastrów danych lub przez przenoszenie danych do przestrzeni o mniejszym

wymiarze, w której jest łatwiejsza wizualizacja analizowanych danych. W literaturze można znaleźć różne aplikacje, które pokazują, jak algorytmy DM mogą być stosowane do odkrywania wiedzy z różnych obszernych baz danych, np: zagadnienie przewidywania niezadowolenia klientów [19, 157], klasteryzacja dokumentów w archiwum [102], śledzenie i identyfikacja tropikalnych cyklonów [141], diagnostyka raka z mammogramów [236]. Zastosowanie ANN w zagadnieniach wydobywania wiedzy przedstawiono w artykule Shine'a [214], gdzie sieć neuronową użyto do generowania wag dla cech danych. Wygenerowane wagi w dalszym etapie wykorzystano do klasteryzacji danych w oparciu o algorytm k-najbliższych sąsiadów.

W literaturze obejmującej zagadnienia zastosowania hurtowni danych w diagnostyce często opisuje się trzy podstawowe zadania stawiane przed algorytmami DM [17]:

* klasteryzacja,
* wizualizacja,
* wydobywanie reguł.

Klasteryzacja i samoorganizujące się mapy (ang. *Self Organizing Map, SOM*) są jednym z wielkich odkryć DM. Zagadnienie to zostało dokładnie opisane w artykule Kohonena [102] obejmując zadanie organizacji i odzyskiwania dokumentów z archiwów. Kohonen zademonstrował słabą użyteczność wielkich samoorganizujących map zawierających ponad jeden milion węzłów, które miały dokonać klasteryzacji nieco mniej niż 7 milionów opisów patentów. Dokumenty te zostały scharakteryzowane za pomocą 500 wymiarowego wektora cech. Dużo lepszą skuteczność można uzyskać dzieląc zbiory danych w obszary. Zadanie to można rozwiązać dokonując obniżenia wymiarowości bazy danych.

Innym podejściem do przetwarzania dużych zbiorów danych jest zastosowanie strategii krokowej. Vesanto w swoim artykule [232] przedstawił dwukrokową strategię klasteryzacji dużych zbiorów danych. W pierwszym kroku dokonuje się podziału danych stosując technikę SOM. Prototypowe rezultaty SOM są w drugim kroku dzielone na klastry za pomocą metody k-średnich (ang. *k-mean*) [118, 143]. SOM z jego sąsiedzkimi ograniczeniami redukuje stopień swobody przygotowując dane do następnego kroku klasteryzacji.

Wizualizacja opiera się często na wstępnym zadaniu redukcji wymiaru danych. Zadanie redukcji wymiaru można zrealizować za pomocą nieliniowych map Sammons'a. König w swoim artykule [136] proponuje połączyć techniki ANN do kombinacji SOM i nieliniowego kojarzenia map Sammons'a do przetwarzania dużych zbiorów danych. W artykule Wanga [236] przedstawiono nowy hierarchiczny algorytm, który pozwala kompletować zbiory danych stosując sieci neuronowe. Metodę

tę zastosowano do badań diagnostycznych, bazując na decyzjach systemu ekspertowego, przy badaniach raka piersi w oparciu o cyfrową reprezentację mammogramu.

Wydobywanie reguł. W wielu zastosowaniach użytkownicy chcą interpretacji wydobytej wiedzy. Fu i Shortliffe [59] skoncentrowali swoje badania na zasadach wydobywania połączeń między zmiennymi w dużej bazie danych. Do realizacji zadania zastosowali model współczynników pewności i ANN. Wyniki uzyskane w predykcji rzeczników w zadaniu biologii molekularnej pokazały dużą odporność przedstawionego algorytmu na zakłócenia. Zhang [249] zaproponował do budowy algorytmu DM zastosowanie kombinacji ANN i logiki rozmytej. Rozmyta lingwistyczna reprezentacja danych jest przetwarzana w numeryczne cechy i ANN końcowo generuje zbiór rozmytych reguł.

Zadanie wydobywania reguł z przykładów nazywane jest często uczeniem z przykładów. Polega to na poszukiwaniu reguł w danej dziedzinie na bazie dostępnych danych (jeśli takie reguły istnieją). Uczenie reguł pojawia się jako ważne zagadnienie zarówno w uczeniu maszyn jak i wydobywaniu wiedzy. Uczenie maszyn postrzegane jest jako poszukiwanie programów komputerowych, które uczą się przetwarzać elementy wiedzy, podczas gdy DM dotyczy poszukiwania wzorców lub reguł ukrytych w danych.

Istnieje wiele technik uczenia reguł. W symbolicznej sztucznej inteligencji uczenie jest często realizowane w formie przeszukiwania zdefiniowanej przestrzeni hipotez. Zatem uczenie reguł można realizować na bazie:

- przeszukiwania heurystycznego [23, 36],
- drzew decyzyjnych [206],
- logicznego indukcyjnego programowania [140],
- sztucznych sieci neuronowych (ANN) [254],
- algorytmów genetycznych.

Heurystyczne metody są często wprowadzane w celu uniknięcia algorytmów o wykładniczym wzroście złożoności obliczeniowej, niestety, jest to realizowane kosztem niekompletnego przeszukiwania. Żadna z metod nie gwarantuje możliwości nauki prawidłowych reguł z ograniczonej liczby obserwowanych danych. To jest szczególnie ważne, gdy przeszukiwana domena jest złożona. Poza tym wykazują one różnice w złożoności obliczeniowej, ponieważ realizują różne kryteria optymalizacji procesu przeszukiwania.

Symboliczne heurystyczne przeszukiwanie. Metoda bazuje na bezgradientowej (np. generuj-testuj) lub opartej na gradiencie funkcji kosztu (np. wspinania się na szczyt lub najpierw najlepszy) technice przeszukiwań zdefiniowanej przestrzeni hipotez.

Kryterium jakości przeszukiwania bazuje na dokładności i zbieżności algorytmu, parametrów niezbędnych dla oceny i wyboru reguł. Słabością tych metod jest problem prawidłowej definicji wspomnianego kryterium jakości. Niestety, często te parametry są źle zdefiniowane, szczególnie gdy występują szumy, niekonsekwencje i niepewności w analizowanych danych. Poza tym ważnym zagadnieniem tej metody jest problem globalnej optymalizacji funkcji jakości przeszukiwań.

Drzewa decyzyjne. Metody oparte o drzewo decyzyjne wstępnie reprezentują wydobytą wiedzę w postaci drzewa. W drugim etapie drzewo jest transformowane w zbiór reguł. Drzewo decyzyjne jest konstruowane przez sekwencyjny wybór atrybutów, bazując na danych pomiarowych. Popularnym narzędziem uczenia reguł metodą przeszukiwania drzewa decyzyjnego jest algorytm $C4.5$ [206]. Wadą wspomnianej metody jest brak kompletnego, wielowariantowego przeszukiwania przestrzeni hipotez.

Odwrócona logiczna dedukcja. W tym podejściu nauczanie przebiega poprzez generowanie hipotezy, która wraz z pewną ilością podstawowej wiedzy (*a priori*) oraz dostępnymi danymi jest weryfikowana. W przypadku pozytywnego efektu weryfikacji hipoteza staje się nową regułą. Przedstawiona technika generowania reguł nie umożliwia generowania prawidłowych reguł w przypadku występowania w danych zakłóceń, niekonsekwencji i niepewności. Przeszukiwanie przez przestrzeń hipotez jest trudne do realizacji w ogólnym przypadku i złożone z dużą ilością wstępnej bazowej wiedzy.

Sztuczne sieci neuronowe. Ostatnio dobre efekty procesu uczenia reguł osiąga się stosując ANN. Wynika to z możliwości realizacji zadania wielowariantowego przeszukiwania i generowania wyników w warunkach niepewności wiedzy. Do istotnych wad tego rozwiązania zalicza się tak zwany efekt *czarnej skrzynki*, gdzie kolejnym nie rozwiązanym zadaniem jest znalezienie powiązania między uzyskanymi parametrami sieci a interfejsem objaśniania powiązań w wydobytej tą metodą wiedzy. Niezbędne niekiedy objaśnienia, przekazywane obsłudze operatorskiej, wymagają utworzenia wiedzy deklaratywnej reprezentowanej w postaci reguł. Stosowane w zadaniu wydobycia reguł z ANN techniki nie mogą wydobyć wszystkich reguł w przypadku dużych sieci neuronowych oraz wydobyte z takich sieci reguły są często przybliżone. W tym podejściu ANN uczą się funkcji dopasowania do przedstawionych danych, aby w następnym etapie, korzystając z zakodowanej funkcji przejścia sieci, zdekodować ją do zbioru reguł. Proces uczenia sieci ze zgromadzonych danych jest dobrze opisany w literaturze, ale pozostaje do rozwiązania zagadnienie prawidłowego wyekstrahowania reguł z nauczonej sieci.

Algorytmy genetyczne. Stosowanie w procesie przeszukiwania algorytmów genetycznych wymaga zdefiniowania dla każdego zbioru reguł kodu bitowego oraz operatorów genetycznych. Stochastyczna natura algorytmu genetycznego łagodzi efekt lokalnego minimum, ale element losowości może także wprowadzić pewien stopień

nieprecyzyjności uzyskanych wyników. Doświadczenia pokazują, że to podejście do procesu przeszukiwań wprowadza błędy w uczeniu reguł nawet w przypadku niezbyt złożonych domen [59].

Wymienione metody pozwalają wnioskować, że wydobywanie wiedzy z dużych baz danych przy zastosowaniu jednej metody bazującej na określonej reprezentacji wiedzy jest często nieskuteczne. Ponadto w pracy [59] ukazano, że do określonych zadań KDD skutecznymi są wybrane techniki z zadanymi reprezentacjami wiedzy.

Integrowanie różnych form reprezentacji wiedzy zostało również wykorzystane w pracy [59], w której zaprezentowano nowy system DOMRUL wydobywania reguł ze zbioru danych. Charakterystyczną cechą systemu jest zintegrowanie wzajemnie uzupełniających się bloków: sieci neuronowej i modelu współczynnika niepewności CF. Omawiany system wydobywania reguł może być stosowany dla dowolnej dziedziny wiedzy bez ograniczeń zarówno na liczbę i wielkość reguł.

Systemy sterowania i nadzorowania procesu często zbierają duże ilości danych, które można wykorzystać do akwizycji nowej wiedzy. Ta wiedza może być odkryta w bazach danych zgromadzonych podczas monitorowania rzeczywistych procesów lub przygotowanych podczas numerycznych eksperymentów z modelami badanych obiektów.

Zagadnienie akwizycji wiedzy w bazie danych obejmuje odkrywanie zarówno ilościowe jak i jakościowe zależności między atrybutami opisującymi diagnozowany obiekt. Wyznaczone w procesie odkrywania funkcjonalne zależności mogą być zastosowane do predykcji wewnętrznych stanów diagnozowanych obiektów [155].

Dane zbierane z rzeczywistych obiektów wykazują cechy rozmytości, zawierają szumy i niekompletną informację o obiekcie, dlatego wydobycie zależności funkcjonalnych nie zawsze jest możliwe. Dla dużych baz danych można badania podzielić na dwa etapy [155]. W pierwszym następuje stopniowe dzielenie przestrzeni atrybutów na plastry, kierując się ujawnionymi regularnościami między sterowaniem a zależnymi atrybutami. W wydzielonym plastrze danych poszukuje się, najlepszych do znalezienia funkcjonalnych zależności, par niezależnych i zależnych atrybutów. Do zrealizowania tego zakresu badań można zastosować tablicę wielodzielczą i V test Cramera. Kolejny etap zawiera dalsze poszukiwania równań wykonane przy użyciu metody Bacon-3 [154], w której dokonując stopniowego uogólniania znalezionych w plastrach zależności, wyznacza się końcowe równania wielu zmiennych.

Innym podejściem do problemu odkrywania wiedzy jest poszukiwanie odwróconych równań, które mogą być odkrywane bezpośrednio z danych, jeśli zamienimy role miedzy niezależnymi i zależnymi atrybutami. W wyniku przeprowadzonych operacji otrzymuje się odwrócony model, pozwalający na wnioskowanie o klasie stanu technicznego na podstawie symptomów diagnostycznych (wnioskowanie o

przyczynie na podstawie *skutku*) [29]. Po wykonaniu operacji zamiany funkcji atrybutów do dalszych badań można zastosować metodologię opisaną wcześniej.

Ogólnie uzyskane wyniki pozwalają wnioskować, że metodologia KDD może być w pełni użyteczna w wydobyciu wiedzy niezbędnej w procesie diagnostycznym. Ponadto pozwala to wydobywać wiedzę, unikając metody *na ślepo*. Stosując procedury KDD do obszernych baz danych, można uzyskać wiedzę obejmującą szerokie obszary, bardziej wartościową, bardziej akceptowalną i zrozumiałą dla człowieka. Te cechy wpływają na możliwość pełnej automatyzacji całego procesu wydobywania wiedzy.

W procesie gromadzenia danych dla potrzeb diagnostycznych występuje problem dużej ilości cech mierzonych sygnałów obiektowych. Z praktycznego punktu widzenia zastosowania systemu ekspertowego wynika, iż liczba obserwowanych cech powinna być ograniczona. Cechą negatywną tej metodologii jest konieczność dopasowywania formy reprezentacji wiedzy do rozwiązywanego problemu. W celu uzyskania efektywnego systemu diagnostycznego proponuje się zastosowanie różnych zintegrowanych form reprezentacji wiedzy.

Wstępne przetwarzanie zgromadzonych w procesie monitorowania danych obejmuje zagadnienia [32]:

- ograniczenia w zbiorze wartości,
- ograniczenia w zbiorze czasu.

Ograniczenie w zbiorze wartości można realizować m.in. metodą klasyfikacji, reprezentacji przybliżonej, reprezentacji rozmytej lub kwantowania. Kwantowanie danych w diagnostycznej bazie pozwala na obniżenie liczby danych opisujących stan procesu. Polega to na wyborze jednej danej reprezentującej grupę danych o zbliżonej wartości. Proces ten przedstawia następująca zależność:

$$x_i(t) = \widehat{x}_i(t)^{\pm 0.5\delta}, \tag{3.4}$$

gdzie: δ oznacza krok kwantowania.
Dla danych o wartościach zbliżonych do wartości granicznych stosuje się kwantowanie z histerezą [32].

Ograniczenia w zbiorze czasu wiążą się z odpowiednim doborem częstotliwości pobierania danych. Po dokonaniu ograniczenia w zbiorze czasu (efekt sklejania) zmodyfikowana reprezentacja wiedzy w formie trójek przyjmie postać:

$$x = \langle\, o,\ n(x),\ val(x),\ [t_p, t_k]) \,\rangle.$$

gdzie: $[t_p, t_k]$ oznacza przedział czasu obserwacji danych o zbliżonej wartości cechy.

Do oceny dopuszczalnego stopnia ograniczenia reprezentacji cech w zbiorze rozróżnialnych wartości można zastosować metody analizy statystycznej (test χ^2).

3.6. Zakończenie

W rozdziale przedstawiono różne formy reprezentacji wiedzy, zarówno deklaratywnej jak proceduralnej. Porównano zasady reprezentowania wiedzy z wykorzystaniem metod symbolicznych i niesymbolicznych. Omówiono również ważne zagadnienia akwizycji wiedzy od eksperta oraz wydobywania wiedzy z baz danych. Wyjaśniono specyfikę akwizycji wiedzy od eksperta, formy weryfikacji tej wiedzy i technik prowadzenia wywiadów oraz ważenia wybranych elementów wiedzy uzyskanej od kilku ekspertów. Przedstawiono metody pozyskiwania wiedzy z baz danych.

Wyjaśniono pojęcie odkrywania wiedzy i wydobywania wiedzy jako ważne elementy w procesie przygotowania bazy wiedzy systemu diagnostycznego. Przedstawiono pojęcie hurtowni danych jako procesu tworzenia baz danych analitycznych zaliczanego do grupy problemów przygotowania i wykorzystania danych operacyjnych w diagnostyce [203]. Umieszczone w tej części monografii przykłady pokazują różne techniki eksploracji wiedzy w bazach danych. Ukazano w nich zasady wykorzystania różnych technik sztucznej inteligencji w przetwarzaniu dużych baz danych.