Notatki robocze nt. algorytmów indukcji reguł (pełna wersja w skrypcie Uczenie maszynowe, Wyd. PP 2004)

Podstawy

Reguły decyzyjne to jedna z najpopularniejszych metod reprezentacji wiedzy stosowanych w uczeniu maszynowym i odkrywaniu wiedzy z danych. Są one złożone z części warunkowej i części decyzyjnej, tzn. podają decyzję właściwą dla sytuacji, w której spełnione są określone warunki. Najczęściej są one reprezentowane w postaci:

jeżeli spełnione określone warunki *to* decyzja.

Wielu autorów uważa, że reprezentacja wiedzy w postaci reguł jest najbliższa metodom stosowanym do zapisu wiedzy przez ludzi (porównaj dyskusje w [BZ92, C00, Mi83, MBK98, Mi97, Ste01, Zyt02]). Uznawana jest za czytelniejszą dla człowieka niż inne reprezentacje. Podkreśla się także jej modularność i przydatność do analizy pojedynczych reguł przez eksperta. Forma ich prezentacji wykorzystywana jest do finalnej reprezentacji innych skutecznych klasyfikatorów, np. drzewa decyzyjnego czy sieci neuronowych. Reguły decyzyjne są wykorzystywane w wielu skutecznych zastosowaniach uczenia maszynowego [LS98, Ste01].

Zgodnie z przeglądem podanym przez Żytkowa [Zyt02] wyróżnia się trzy podstawowe rodzaje reguł wygenerowanych z danych, tj. **klasyfikacyjne**, **charakterystyczne** oraz **asocjacyjne**.

Pierwszy typ reguł jest ściśle związany z „klasycznym" indukcyjnym uczeniem się pojęć (klas) i określa, czy obiekt (przykład) spełniający warunki elementarne z części warunkowej reguły należy do klasy decyzyjnej (pojęcia) wskazywanego przez jej część decyzyjną. Reguły charakterystyczne (ang. *characteristic rules*) wskazują najbardziej charakterystyczne właściwości obiektów należących do pewnej klasy. Natomiast reguły asocjacyjne (ang. *association rules*) reprezentują wiedzę o tym, czy pewne wartości niektórych atrybutów odpowiednio często współwystępują z pewnymi wartościami innych atrybutów.

Można tutaj wspomnieć o jeszcze innym rodzaju reguł będących formułami logicznymi w postaci klauzul Horna, wykorzystywanych w ramach tzw. **indukcyjnego programowania logicznego** – rozszerza się tam definicje pojęcia i danych na relacje zachodzące między obiektami. Czytelnik może zapoznać się z przeglądami tego typu reguł przedstawionymi w rozdziale 9 książki [C00] lub w rozdziale 10 pozycji [Mi97] czy w rozdziale 3 w [MBK98].

Reguły klasyfikacyjne są najczęściej rozważane zarówno w badaniach, jak i w zastosowaniach praktycznych. W dalszej części rozdziału będzie omówiony wyłącznie ten typ reguł, nazywanych także **regułami decyzyjnymi**. Reguły są indukowane na podstawie uogólnienia zbioru przykładów uczących opisanych za pomocą zbioru atrybutów. Reguły wygenerowane dla każdej klasy decyzyjnej powinny być spełniane przez przykłady należące do tej klasy (tzw. **przykłady pozytywne**). Równocześnie reguły nie powinny być spełniane przez żaden przykład z innych klas lub powinny być spełniane tylko przez niewiele z nich (tzw. **przykłady negatywne**).

Indukcję reguł można rozpatrywać w dwóch perspektywach: **predykcji klasyfikacji** lub **opisu**. W pierwszej perspektywie celem jest wygenerowanie z danych zbioru reguł, które będą użyte do klasyfikowania nowych obiektów. W przypadku perspektywy opisu celem jest odkrycie z danych reguł reprezentujących regularności, które charakteryzują te dane i są „interesujące" dla różnego rodzaju użytkowników. Tego typu wzorce mogą ułatwić interpretacje zależności pomiędzy wartościami atrybutów a definicją klasy decyzyjnej [Ste01].

W rozdziałach 2 i 3 pokazano, że zbiór reguł może być otrzymany przez przekształcenia reprezentacji drzewa decyzyjnego. Bardziej naturalnym sposobem postępowania jest jednak bezpośrednia indukcja reguł z danych. Dlatego w tym rozdziale będzie omawiana podstawowa strategia bezpośredniej indukcji minimalnego zbioru reguł, tj. strategia generowania kolejnych pokryć wykorzystywana w najpopularniejszych algorytmach. Ponadto

dokonamy przeglądu podstawowych algorytmów i szczegółowo zaprezentujemy algorytm o nazwie LEM2. Dalsza część rozdziału zawiera krótką charakterystykę miar oceny reguł oraz opis wykorzystania zbioru reguł do klasyfikowania nowych obiektów. Celem ćwiczenia jest zapoznanie się z wybranymi algorytmami indukcji reguł, strategiami klasyfikowania za ich pomocą oraz porównanie tego typu podejść z innymi klasyfikatorami.

Reprezentacja reguł

Powtórzmy założenie przedstawione w rozdziale 1, zgodnie z którym przykłady uczące reprezentowane są w postaci tablicy informacyjnej (U,A) lub tablicy decyzyjnej $(U,A\cup\{d\})$, gdzie U jest niepustym zbiorem obiektów (przykładów), A jest niepustym i skończonym zbiorem atrybutów, V_a jest dziedziną atrybutu $a\in A$. Na zbiorze obiektów zdefiniowany jest zbiór klas decyzyjnych $K = \{K_j : j = 1,..., r\}$. Określenie pojęć może być podane przez „nauczyciela" lub zdefiniowane za pomocą dodatkowego atrybutu d opisującego obiekty w tablicy decyzyjnej.

Indukcja reguł decyzyjnych dotyczy zbioru klas (pojęć) K. Zbiór obiektów dzielony jest na podzbiory $E_{K1}\cup E_{K2}\cup\ldots\cup E_{Kr}$ zawierające przykłady z poszczególnych klas. Niech $U = E^+_{Kj} \cup E^-_{Kj}$ dla danego pojęcia K_j oznacza podział na podzbiór przykładów pozytywnych $E^+_{Kj} = E_{Kj}$ oraz przykładów negatywnych $E^-_{Kj} = S \setminus E_{Kj}$ (tj. pozostałych przykładów).

Reguła decyzyjna r opisująca pojęcie K_j zdefiniowana jest jako wyrażenie postaci:

$$\text{jeżeli } P \text{ to } Q, \qquad (4.1)$$

gdzie P jest częścią warunkową (przesłanką) reguły oraz Q jest częścią decyzyjną (konkluzją) reguły oznaczającą, że obiekt x spełniający P należy do K_j.

Część warunkowa P jest koniunkcją warunków elementarnych w i jest reprezentowana w postaci: $P = w_1\wedge w_2\wedge\ldots\wedge w_k$, gdzie k jest liczbą wykorzystanych warunków. Niektórzy autorzy koniunkcję P nazywają złożeniem warunków lub kompleksem [C00, Mi83].

Warunek elementarny w_i dla danej reguły jest zdefiniowany jako wyrażenie $(a \propto v_a)$, gdzie $\propto$ oznacza operator relacji, najczęściej ze zbioru $\{=,\neq,<,\leq,>,\geq\}$, a v_a oznacza stałą będącą wartością z dziedziny atrybutu a. Możliwe jest też użycie warunku $(a \in G_a)$, gdzie G_a jest podzbiorem wartości z dziedziny atrybutu a[1]. Należy zauważyć, że w powyższym zapisie warunku elementarnego atrybut a przybiera wybraną wartość, np. może to być zapis (*dochód* = niski). W takiej notacji atrybut traktowany jest jak zmienna. W celu większej jednoznaczności notacji część autorów stosuje zapis $(a(x) \propto v_a)$, gdzie $a(x)$ oznacza wartość atrybutu a dla obiektu x.

Aby zilustrować notację reguł, przedstawmy zapis reguły, którą można wygenerować ze zbioru przykładów przedstawionego w tabeli 1.1:

Jeżeli (*dochody* ≤ 1500 zł)∧(*cel_kredytu* = samochód) to (*kredyt* = ryzyko).

Wprowadzimy poniżej kilka definicji przydatnych dla przedstawienia algorytmów indukcji reguł.

Pokryciem (ang. *cover*) koniunkcji warunków elementarnych P, oznaczonym przez $[P]$, jest zbiór obiektów spełniających logicznie warunki elementarne reprezentowane przez P. Pokrycie $[P]$ można podzielić na część pozytywną $[P]^+_{Kj} = [P]\cap E^+_{Kj}$ oraz negatywną $[P]^-_{Kj} = [P]\cap E^-_{Kj}$.

Reguła decyzyjna r jest **dyskryminująca**, tj. odróżnia przykłady pozytywne należące do K_j od przykładów negatywnych wtedy i tylko wtedy, gdy jej część warunkowa spełnia warunek niesprzeczności $[P]^-_{Kj}=\varnothing$. Wymaga się także, aby $[P]^+_{Kj}\neq\varnothing$. Dyskryminująca reguła

[1] W literaturze można spotkać także inny sposób zapisu tego typu warunku, tj. jako wyrażenie $(a = v_f\vee v_{f+1}\vee\ldots\vee v_h)$, gdzie $v_f, v_{f+1},\ldots, v_h$ są wartościami z podzbioru G_a.

decyzyjna r jest **minimalna**, jeśli usunięcie jakiegokolwiek warunku elementarnego w z jej części warunkowej P prowadzi do sytuacji $[P]^{-}_{Kj} \neq \emptyset$.

Zbiór reguł decyzyjnych R opisuje w pełni pojęcie K_j, jeśli każdy przykład $x \in E^{+}_{Kj}$ spełnia część warunkową i decyzyjną przynajmniej jednej reguły $r \in R$. Michalski [Mi83] podał, że tworzenie regułowego opisu klas może przybierać postać tzw. **opisu dyskryminującego**[2], który musi spełniać wymóg pełnego opisu oraz zawierać dyskryminujące i minimalne reguły. Pojęcie minimalnego opisu dyskryminującego oznacza, że zbiór reguł jest minimalny, tzn. nie istnieje żaden jego podzbiór, który spełnia warunki kompletności i niesprzeczności.

Algorytmy indukcji reguł

Strategie sekwencyjnego pokrywania

Najbardziej znane algorytmy opierają się na zasadzie **generowania kolejnych pokryć** (ang. *sequential covering*). Polega ona na uczeniu się pojedynczej reguły, usuwaniu przykładów, które ona pokrywa i powtarzaniu procesu dla pozostałych przykładów. W rezultacie powstaje zbiór reguł pokrywających rozważany zbiór przykładów. Najczęściej algorytmy są tak skonstruowane, aby iteracyjnie powtarzać strategie sekwencyjnego pokrywania dla przykładów z kolejnych klas. Poniżej podano ogólny schemat takiego algorytmu dla pojedynczej klasy K_j:

Procedure Sequential covering (K_j – klasa; A – atrybuty; E – przykłady, τ – próg akceptacji reguły);
begin
  $R := \emptyset$; {zbiór poszukiwanych reguł}
  $r :=$ ***learn-one-rule***(klasa K_j; atrybuty A; przykłady E)
  while ***ocena***$(r,E) > \tau$ **do**
  begin
    $R := R \cup r$;
    $E := E \setminus [R]$; {usuń przykłady pozytywne pokryte przez R}
    $r :=$ ***learn-one-rule***(klasa K_j; atrybuty A; przykłady E);
  end;
return R
end.

Przy zapisie algorytmu założono, że dostępna jest funkcja ***learn-one-rule***, która dla danego zbioru przykładów znajduje jedną regułę pokrywającą możliwie jak najwięcej przykładów pozytywnych i jak najmniej negatywnych. Funkcja może być zrealizowana różnymi sposobami w zależności od konkretnego algorytmu (porównaj dyskusję w [Mi97, Kl02, Ste01]). W wielu z tych algorytmów początkowy „kandydat" na część warunkową reguły pokrywa zbiór wszystkich przykładów, w tym negatywnych. Następnie podlega specjalizacji poprzez dodawanie warunków elementarnych, dopóki nie zostanie spełniony warunek akceptacji reguły (np. niesprzeczności w algorytmie AQ). ***Ocena*** jest zależną od użytkownika funkcją oszacowania jakości zbioru reguł. Na przykład kolejnych reguł poszukuje się, dopóki pozostają przykłady pozytywne z klasy K_j nie pokryte przez żadną z dotychczas wygenerowanych reguł [Mi83]. Można też za pomocą progu τ zdefiniować warunek wcześniejszego zakończenia poszukiwania zbioru reguł [CN2]. Algorytm powtarza się iteracyjnie oddzielnie dla każdej klasy K_j lub w niektórych algorytmach dla wszystkich klas. W drugim przypadku należy odpowiednio rozróżniać przykłady pozytywne [CN2, Ste01]. Należy zauważyć, że przedstawiony schemat wykorzystuje heurystykę „zachłanną", gdyż usuwa przykłady pokryte przez wygenerowaną regułę. Podobnie heurystycznie ogranicza się przestrzeń przeszukiwania warunków elementarnych w trakcie budowy pojedynczej reguły.

[2] Pojęcie K_j jest określane także jako tzw. lokalne pokrycie (ang. *local covering*).

Na ogół generowany zbiór reguł jest nieuporządkowany, choć niektóre algorytmy pozwalają także przedstawić zbiór reguł w postaci listy uporządkowanej – wtedy reguły są ułożone zgodnie z ich przydatnością do klasyfikowania nowych przypadków.

Do oceny „kandydatów" na części warunkowe reguł stosuje się różne miary w zależności od algorytmu. Michalski [Mi83] proponuje rozważać szereg kryteriów, które ustawione są w porządku leksykograficznym. W [MBK] jako najczęściej stosowane miary wymienia się:

- maksymalizację liczby przykładów pozytywnych pokrywanych przez koniunkcję P, tj. $|[P]^{+}_{Kj}|$, gdzie $|.|$ oznacza liczność zbioru,
- maksymalizację stosunku liczby pokrywanych przykładów pozytywnych do ogólnej liczby wszystkich przykładów, tj. $|[P]^{+}_{Kj}|/|[P]|$,
- minimalizację liczby użytych warunków elementarnych.

Inne algorytmy, np. CN2 [CN2] czy MODLEM [Ste97], wykorzystują do oceny koniunkcji warunków P miarę entropii informacji w podobny sposób jak w indukcji drzew decyzyjnych. Jeszcze inną miarą oceniającą zależność wyrażeń P i Q jest estymata Laplace'a będąca szczególną m-estymatą prawdopodobieństwa:

$$Lp(P) = \frac{n_c + lr - 1}{n + lr}, \qquad (4.2)$$

gdzie n_c jest liczbą przykładów pozytywnych pokrywanych przez koniunkcję P, n – ogólną liczbą wszystkich przykładów pokrywanych przez koniunkcję P, a lr jest liczbą klas decyzyjnych.

W podejmowaniu decyzji o wyborze najlepszych warunków elementarnych podczas indukcji reguł można także korzystać z zasady minimalnej długości opisu (ang. MDL – *Minimal Description Length*) [C00, Mi97]. Zasada ta jest związana z kodowaniem danych oraz reprezentacji hipotez (patrz przykłady podane w [C00]). W łącznym kodzie wyróżnia się część związaną z dopasowaniem hipotezy do przykładów uczących (interpretowaną jako błąd na zbiorze uczącym) oraz część związaną ze złożonością hipotezy (rozumianej jako długość jej optymalnego kodu). Zgodnie z zasadą minimalności opisu dla danego zbioru przykładów i różnych hipotez należy wybrać hipotezę, która minimalizuje łączną długość kodu.

Historycznie pierwszym algorytmem opartym na zasadzie pokryć był algorytm AQ podany przez Michalskiego. Jego kluczowym operatorem budowy hipotez była tzw. gwiazda (ang. *STAR*) specjalizująca „maksymalnie ogólne" warunki pokrywające wybrany początkowy przykład pozytywny (tzw. zalążek) w celu odróżnienia ich od przykładów negatywnych. W pierwotnej postaci AQ tworzył minimalne opisy dyskryminujące pokrywające wszystkie przykłady (z bardziej szczegółowym opisem algorytmu czytelnik może się zapoznać np. w [BZ92, C00, MBK]). Podejście to było rozwijane w rodzinie algorytmów AQ11, AQ15 i AQ17 oraz w systemie INLEN. Algorytm CN2 [CN2] łączy inspirację ideą algorytmu AQ z technikami znanymi z indukcji drzew decyzyjnych w celu lepszego uwzględniania zaszumionych przykładów i unikania zjawiska przeuczenia. Stosuje także inny sposób tworzenia hipotez: od ogólnej do szczegółowej. Algorytm ten jest wykorzystywany w systemie odkrywania wiedzy MLC++.

Inne znane algorytmy to Itrule, korzystający z zasad wnioskowania probabilistycznego, PRISM zgodny z zasadą kolejnych pokryć, lecz w inny sposób poszukujący najlepszej z reguł, MODLEM, przeznaczony do uwzględniania podczas indukcji także atrybutów ilościowych, PVM, poszukujący najbardziej predyktywnych warunków elementarnych, czy HCV, wykorzystujący tzw. macierze rozszerzeń (patrz przegląd w [Ste01]). W [Mi97] omówiono sposób poszukiwania reguł za pomocą algorytmów genetycznych. Należy ponadto wspomnieć o podejściach do generacji zbiorów reguł innych niż minimalny. Podejścia te są ukierunkowane na poszukiwanie pełniejszego podzbioru wszystkich reguł w przyjętej składni, które można wygenerować z danego zbioru przykładów. W tej grupie podejść za szczególnie interesującą należy uznać metodę **wnioskowania boolowskiego** opartą na tzw.

funkcjach rozróżnialności, zaproponowaną przez Skowrona [Sk93][3]. Pełniejszą dyskusję różnych propozycji czytelnik może znaleźć np. w [Kl02, Ste01].

Algorytm LEM2

W celu zilustrowania algorytmów indukcji reguł wykorzystujących strategię kolejnych pokryć zaprezentujemy algorytm LEM2 zaproponowany przez Jerzego Grzymałę-Bussego w [G92]. Omówimy jego podstawową wersję przedstawioną w zapisie dla spójnych przykładów oraz reprezentacji warunków elementarnych w postaci $(a = v_a)$. W przypadku występowania sprzecznych przykładów w algorytmie wykorzystano teorię zbiorów przybliżonych (ang. *rough sets*) wprowadzoną przez Pawlaka [Pa91][4]. Istnieją także bardziej zaawansowane wersje tego algorytmu, to jest LEM2 – *with-interval-extension* lub MLEM [G96, G02].

Algorytm LEM2 wykorzystuje heurystyczną strategię generującą minimalny zbiór reguł. Jest ona uruchamiana iteracyjnie dla kolejnych klas decyzyjnych (lub ich przybliżeń). W dalszej prezentacji skupimy się na pojedynczej klasie decyzyjnej K_j. Przez K oznaczać będziemy zbiór obiektów E_{Kj} będących przykładami pozytywnymi wybranej klasy.

W przypadku algorytmu LEM2 warunek elementarny w dla atrybutu $a \in A$ jest reprezentowany jako wyrażenie $(a = v_a)$[5], gdzie v_a jest wartością z dziedziny V_a. Ponadto w zapisie będzie się stosować notacje: $[w]$ oznacza pokrycie warunku elementarnego, czyli zbiór przykładów spełniających wyrażenie w; $W(G)$ oznacza zbiór różnych warunków elementarnych w ustalonej składni, które można skonstruować na podstawie opisów przykładów ze zbioru $G \subseteq K$ (przy przyjętej składni warunki budowane są na podstawie kolejnych wartości atrybutów występujących w opisie przykładów z G). $P = w_1 \wedge w_2 \wedge \ldots \wedge w_m$ jest koniunkcją warunków elementarnych, która jest „kandydatem" na część warunkową reguł. Koniunkcja P może być zaakceptowana jako część warunkowa reguły wskazującej klasę K_j, jeżeli:

$$\varnothing \neq [P] = \bigcap_{i=1}^{m} [w_i] \subseteq K . \qquad (4.3)$$

Ponadto P powinno być minimalną koniunkcją spełniającą warunek (4.3), tzn. usunięcie któregokolwiek z warunków w_i prowadzi do $[P] \not\subset K$.

W stosowanej notacji **P** jest zbiorem koniunkcji P stanowiącym tzw. **lokalne pokrycie** zbioru K (definicja w rozdziale 4.2). Wymaga się, iteracji wykonywanej dla zbioru obiektów K przedstawiony jest poniżej.

Procedure LEM2
(**input:** K –zbiór przykładów pozytywnych pojęcia K_j;
output: R – zbiór reguł opisujących pojęcie K_i);
begin
  $G := K$; {zbiór obiektów nie pokrytych dotychczas przez elementy z **P**}
  P := ∅; {lokalne pokrycie **P** zbioru K}
  while $G \neq \varnothing$ **do**
  begin
    $P := \varnothing$; {kandydat na część warunkową reguły}
    $W(G) := \{w : [w] \cap G \neq \varnothing\}$; {zbiór potencjalnych warunków elementarnych}
    while $(P = \varnothing)$ **or not** $([P] \subseteq K)$ **do**
    begin
      wybierz warunek $w \in W(G)$ taki, że wyrażenie $|[w] \cap G|$ ma największą wartość,
      jeśli więcej niż jeden z warunków maksymalizuje powyższe wyrażenie,

[3] Metody wnioskowania boolowskiego są często stosowane w ramach odkrywania wiedzy z niespójnych tablic decyzyjnych opartego na teorii zbiorów przybliżonych [Sk91].

[4] W przypadku niemożliwości precyzyjnego zdefiniowania zbioru obiektów (klasy decyzyjnej) w teorii zbiorów przybliżonych tworzy się dolne i górne przybliżenie tego zbioru na podstawie klas relacji nierozróżnialności pomiędzy obiektami. Indukcję reguł prowadzi się z przykładów należących do przybliżeń klas decyzyjnych (patrz [G92, Pa91, SSGM00, Ste01]).

[5] Ogólniejsze reprezentacje warunków elementarnych, np. $(a \geq v_a)$, rozważa się w algorytmach takich jak MODLEM albo MLEM.

```
        to wybierz ten, dla którego wyrażenie |[w]| ma minimalną wartość,
        w przypadku niejednoznaczności wybierz pierwszy z rozważanych warunków;
        P := P ∪ {w};     {dołącz najlepszy warunek do koniunkcji P}
        G := G ∩ [w];    {ogranicz zbiór obiektów dla stworzenia warunków elementarnych}
        W(G) := {w : [w]∩G ≠ ∅};        {uaktualnij listę potencjalnych warunków}
        W(G) := W(G) – {P};
      end; {while not ([P] ⊆ K)}
      for każdy warunek elementarny w ∈ P do
            if [P – {w}] ⊆ K then P := P – {w};     {usuń nadmiarowe warunki}
      P := P ∪ {P};
      G := K – ∪ P∈P[P];
    end; {while G ≠ ∅}
    for każdy P ∈ P do
      if ∪ T∈P–{P} [T] = K then P := P – {P};      {usuń ewentualne nadmiarowe reguły}
    utwórz zbiór reguł R na podstawie koniunkcji z P;
end {procedure}.
```

W celu ilustracji działania algorytmu LEM2 rozważmy zbiór przykładów uczących zaprezentowanych w tabeli 2.1. W wyniku użycia algorytmu otrzymuje się następujący zbiór reguł (w poniższej notacji po zapisie składni reguły podaje się zbiór przykładów pokrywanych przez każdą z reguł):

jeżeli (*dochody* = średnie) ∧ (*student* = tak) to (*kupuje_komputer* = tak) {1,6}
jeżeli (*dochody* = wysokie) to (*kupuje_komputer* = tak) {3}
jeżeli (*student* = nie) to (*kupuje_komputer* = nie) {2,7,8}
jeżeli (*dochody* = niskie) to (*kupuje_komputer* = nie) {4,5,7}

Należy zwrócić uwagę, że zbiór reguł opisuje wszystkie przykłady obu klas decyzyjnych. Ponadto jest to minimalny zbiór reguł.

Wybrane miary oceny reguł

Reguły mogą być różnie oceniane w zależności od tego, czy są stosowane do predykcji klasyfikacji nowych obiektów czy do budowania opisu zależności w zbiorze uczącym. W perspektywie klasyfikowania ocena dotyczy całego zbioru reguł i jest oparta najczęściej na kryterium trafności klasyfikowania. Wykorzystanie zbioru reguł jako klasyfikatora będzie omówione w podrozdziale 4.4. W drugiej perspektywie opisu reguły traktowane są jako reprezentacja regularności, wzorców (lub czasami anomalii), które są charakterystyczne dla zbioru przykładów uczących. Mogą one ułatwić interpretacje zależności między wartościami atrybutów a wartością klasy decyzyjnej [Ste01]. W literaturze można spotkać wiele propozycji miar oceniających ilościowe właściwości pojedynczych reguł (interesujący przegląd zawarto w [YZ99]).

Większość miar definiuje się na podstawie zbioru przykładów uczących, z których wygenerowano regułę. Podstawowe miary to wsparcie i dokładność reguły. Przypomnijmy, że reguła r reprezentowana jest w postaci: *jeżeli P to Q*. Wtedy w zbiorze n przykładów uczących n_{PQ} oznacza liczbę obiektów spełniających zarówno Q, jak i P, n_P – liczbę obiektów spełniających P, a n_Q oznacza liczbę obiektów spełniających Q. Przy tej notacji **wsparcie reguły** (ang. *support*) jest zdefiniowane jako:

$$\frac{n_{PQ}}{n}. \tag{4.4}$$

Dokładność reguły definiuje się jako:

$$AS(Q \mid P) = \frac{n_{PQ}}{n_P}. \tag{4.5}$$

Miara ma zakres zmienności $0 \le AS(Q|P) \le 1$ i określa stopień, w jakim przesłanka P implikuje konkluzję Q. Można ją interpretować jako prawdopodobieństwo warunkowe, że

losowo wybrany obiekt, który spełnia wyrażenie P, spełnia również Q. Miara (4.5) jest często używana w eksploracji danych, np. do indukcji reguł asocjacyjnych, jako **wiarygodność** (ang. *confidence*). Pawlak pisze w [Pa99], że historycznie ta miara po raz pierwszy została wprowadzona przez Łukasiewicza jako współczynnik pewności implikacji logicznej. Z przeglądu literatury wynika także, że jest ona często wykorzystywana przez innych autorów, choć nadają jej inne nazwy, np. bezwzględne wsparcie reguły, stopień dyskryminacji.

Inną popularną miarą jest **względne pokrycie reguły** (ang. *coverage*) zdefiniowane w następującej postaci:

$$AS(P \mid Q) = \frac{n_{PQ}}{n_Q}. \qquad (4.6)$$

Reguła ma wysokie pokrycie, jeśli obiekty spełniające konkluzję reguły Q równocześnie spełniają wyrażenie P. Może być także interpretowana w kategoriach prawdopodobieństwa warunkowego.

Aby zilustrować obliczanie powyższych miar, rozważmy regułę r o postaci: „jeżeli (*student* = nie) to (*kupuje_komputer* = nie)" wygenerowaną z przykładów uczących reprezentowanych w tabeli 2.1. Pokrywa ona trzy przykłady o identyfikatorach 2, 7, 8, czyli $n_{PQ} = 3$. Równocześnie $n_P = 3$. Zbiór uczący zawiera $n = 8$ przykładów, w tym w klasie decyzyjnej (*kupuje_komputer* = nie) 5 przykładów. Wsparcie reguły r jest równe 3/8, to jest 0,375. Dokładność reguły $AS(Q|P)$ jest równa $3/3 = 1$. Natomiast względne pokrycie tej reguły jest równe 3/5, czyli 0,6.

Klasyfikowanie obiektów za pomocą reguł

Reguły decyzyjne wygenerowane z przykładów uczących używane są do klasyfikowania nowych obiektów (lub przykładów testowych). Klasyfikowanie obiektów opiera się na **dopasowaniu** (ang. *matching*) opisu obiektu do części warunkowych reguł decyzyjnych. Wyróżniamy dopasowanie pełne i częściowe. W przypadku **pełnego dopasowania** (ang. *complete matching*) opis klasyfikowanego obiektu spełnia wszystkie warunki elementarne występujące w części warunkowej reguły. Mówimy o **częściowym dopasowaniu** (ang. *partial matching*) obiektu do części warunkowej reguły, jeśli istnieje przynajmniej jeden warunek elementarny, który nie jest spełniony przez opis klasyfikowanego obiektu[6].

Dopasowanie reguł decyzyjnych do opisu klasyfikowanego przykładu wykorzystywane jest różnie w zależności od tego, czy reguły decyzyjne uporządkowane są w postaci listy decyzyjnej albo czy tworzą nieuporządkowany zbiór reguł.

W przypadku listy decyzyjnej dokonuje się dopasowania obiektu do kolejnych reguł na liście [C4.5]. Pierwsza dopasowana reguła na liście wyznacza przydział klasyfikowanego obiektu do klasy decyzyjnej. Nie przegląda się wtedy dalszych reguł. Ostatnią regułą na liście jest tzw. **reguła domyślna** (ang. *default rule*), stosowana wtedy, gdy żadna z poprzednich reguł nie dopasowała się do obiektu. Reguła domyślna najczęściej wskazuje klasę większościową, tj. najliczniejszą w zbiorze uczącym. Innym rozwiązaniem jest wymuszenie sytuacji, gdy części warunkowe reguł tworzą rozłączny podział przestrzeni przykładów, który prowadzi do jednoznacznych klasyfikacji.

W przypadku klasyfikowania obiektu, oznaczonego dalej przez e, za pomocą nieuporządkowanego zbioru reguł występuje jedna z następujących sytuacji:

1. Część warunkowa dokładnie jednej reguły jest w pełni dopasowana do obiektu, który jest zaklasyfikowany do klasy wskazywanej przez tę regułę.
2. Część warunkowa więcej niż jednej reguły jest w pełni dopasowana do obiektu.
3. Część warunkowa żadnej reguły nie jest dopasowana do obiektu.

Sytuacja (1), a także sytuacja (2), są jednoznaczne, jeśli wszystkie reguły wskazują tę samą klasę, natomiast pozostałe sytuacje rozwiązuje się różnymi sposobami (porównaj przegląd w [G94, Ste01]). Poniżej opiszemy krótko dwie techniki, podane odpowiednio przez Grzymałę [G94] oraz Michalskiego [Mi86].

[6] Często oczekuje się, aby choć jeden z warunków w składni reguł był spełniony.

Pierwszą z tych technik można rozwiązać obie sytuacje (2 i 3) na podstawie dodatkowych miar charakteryzujących regułę r. Grzymała wprowadza współczynnik siły reguły $MR(r) = |[P \wedge Q]|$. Jest to łączna liczba przykładów uczących wspierających regułę. Rozważa także tzw. specyficzność reguły $SR(r)$, zdefiniowaną jako liczba warunków elementarnych w regule.

W sytuacji (2) dla reguł, których części warunkowe są całkowicie dopasowane do obiektu, określa się tzw. poparcie klasy decyzyjnej K_k, zdefiniowane jako:

$$SUP(K_k) = \sum_{r \in R_k^{=}} MR(r) \cdot SR(r), \qquad (4.7)$$

gdzie $R_k^{=}$ oznacza reguły z R_k (wskazujące klasę K_k) w pełni dopasowane do obiektu e. Obiekt jest przydzielany do tej klasy decyzyjnej K_k, dla której poparcie $SUP(K_k)$ osiąga największą wartość.

Rozwiązanie sytuacji (3) polega na zidentyfikowaniu reguł częściowo dopasowanych do obiektu. Dla przypadku częściowego dopasowania określa się dodatkową miarę stopnia dopasowania tej reguły do obiektu $MF(r,e)$. Jest ona równa stosunkowi liczby warunków reguły r dokładnie dopasowanych do opisu obiektu do ogólnej liczby warunków tworzących część warunkową reguły. Przykład e jest klasyfikowany do tej klasy K_k, dla której wyrażenie (4.8) przybiera największą wartość:

$$SUPP(K_k) = \sum_{r \in R_k^{=}} MF(r,e) \cdot MR(r) \cdot SR(r), \qquad (4.8)$$

gdzie $R_k^{\approx}$ oznacza reguły z R_k częściowo dopasowane do obiektu e.

Michalski proponuje w [Mi86] oszacowanie prawdopodobieństwa przydziału obiektu e do klas decyzyjnych (opis tego podejścia dostępny jest także w [BZ92]). W przypadku niejednoznacznej sytuacji (2) dla każdej z reguł r o postaci: *jeżeli P to Q* należących do zbioru R_k oszacowuje się prawdopodobieństwo EP (ang. *estimated probability*) zdefiniowane następująco:

$$EP(r,e) = \begin{cases} |[P \wedge Q]| / n & \text{jeśli } P \text{ jest spełnione przez } e \\ 0 & \text{w przeciwnym przypadku,} \end{cases} \qquad (4.9)$$

gdzie n jest liczbą wszystkich przykładów uczących. Jak widać, wyrażenie (4.9) jest obliczone wyłącznie dla reguł r w pełni dopasowanych do opisu klasyfikowanego obiektu e. Jeśli dla klasy decyzyjnej K_k więcej niż jedna reguła jest w pełni dopasowana, to obliczamy sumę probabilistyczną $EP(K_k,e)$ dla wszystkich tych reguł $r \in R_k^{=}$. Na przykład dla dwóch reguł $r_1, r_2 \in R_k^{=}$ jest to następujące wyrażenie: $EP(K_k,e) = EP(r_1,e) + EP(r_2,e) - EP(r_1,e) \cdot EP(r_2,e)$. Obiekt e jest ostatecznie przydzielany do tej klasy decyzyjnej, dla której wyrażenie $EP(K_k,e)$ osiąga największą wartość.

W sytuacji braku pełnego dopasowania (3) poszukuje się reguł częściowo dopasowanych i oblicza się dodatkowo tzw. miarę dopasowania MF dla każdej klasy w następujący sposób:

- dla warunku elementarnego w_j[7] wartość $MF(w_j,e)$ jest równa 1, jeśli w_j jest spełniony przez e albo w przeciwnym przypadku jest równa wartości ilorazu *l.wartości* / (rozmiar dziedziny atrybutu), gdzie *l.wartości* jest liczbą wartości połączonych wewnętrzną alternatywą[8] w warunku w_j;
- dla reguły r zawierającej k warunków elementarnych w_j jest równa

$$MF(r,e) = \prod_{j=1}^{k} MF(r_j,e) \cdot |[P \wedge Q]| / n$$

[7] Przy założeniu, że atrybut występujący w warunku przybiera wartości ze skończonej dziedziny.

[8] Przypominamy, że w algorytmie AQ warunki elementarne mogą zawierać dysjunkcje kilku wartości, tj. być reprezentowane jako wyrażenie $(a = v_f \vee v_{f+1} \vee \ldots \vee v_h)$, gdzie $v_f, v_{f+1}, \ldots, v_h$ są wartościami z podzbioru V_a.

Ostatecznie dla każdej z klas decyzyjnych K_k oblicza się sumę probabilistyczną $MF(K_k,e)$. Obiekt e jest ostatecznie przydzielany do tej klasy decyzyjnej, dla której ta suma osiąga największą wartość.