ISSN 1232-3829
tts
10/2013
Technika Transportu Szynowego koleje ▪ tramwaje ▪ metro

TROJCZAK-GOLONKA Patrycja, MILEWSKI Stanisław

ZASADY AUTOMATYCZNEGO POZYSKIWANIA REGUŁOWYCH BAZ WIEDZY DLA SYSTEMÓW EKSPERTOWYCH

Streszczenie

W artykule zaprezentowano zalety i wady regułowej reprezentacji bazy wiedzy oraz podstawowe klasy metod jej indukcji maszynowej. Wskazano problemy towarzyszące temu procesowi jak również pokazano na przykładach praktyki i obszary zastosowań regułowych baz wiedzy.

WSTĘP

Najlepiej znaną metodą pozyskiwania danych dla systemów ekspertowych są szczegółowe wywiady z ekspertami poparte obserwacjami zachowań rzeczywistych obiektów, uzupełnione analizą literatury źródłowej. W wielu sytuacjach możliwe jest również dokonanie eksperymentów badawczych, których celem jest zbadanie zasad zachowania rzeczywistych obiektów, których działanie ma być naśladowane przez systemy ekspertowe. Metody te są jednak czasochłonne i zwykle nie gwarantują pozyskania kompletnej wiedzy regułowej, gdyż zwykle wymagają rozstrzygania bieżących sprzeczności logicznych oraz nie zapewniają pokrycia pełnej przestrzeni decyzyjnej.

Gdy nie ma możliwości pozyskania wiedzy regułowej od rzeczywistego eksperta np. z powodu jego braku lub też badania eksperymentalne nie mogą być przeprowadzone z uwagi na ograniczenia zewnętrzne (np. aspekt bezpieczeństwa, aspekt ekonomiczny) wówczas zastosowanie znajduje uczenie maszynowe na podstawie danych przykładowych. Przykłady dostarczane są przez nauczyciela, który może być utożsamiany z modelem lub symulatorem badanego obiektu. Przykładami mogą być też dane pomiarowe rzeczywistego obiektu, jeśli tylko możliwe jest zebranie danych, wcześniej n przez nauczyciela.

1. REGUŁOWA REPREZENTACJA WIEDZY

Wiedza może być reprezentowana i przechowywana na wiele sposobów, przy czym przez reprezentację wiedzy rozumie się sposób, w jaki wiedza o zjawisku, czy pojęciu jest przedstawiana wraz z metodami przetwarzania oraz wnioskowania (inferencji). Siłą sprawczą determinującą zakres i kierunek prac nad reprezentowaniem wiedzy jest to, do czego owa reprezentacja ma być stosowana oraz – w pewnym stopniu – to w jaki sposób wiedza będzie pozyskiwana.

Oprócz języka naturalnego najbardziej zrozumiałą dla człowieka formą opisu są reguły logiczne, które mogą być zapisane w różnych językach logiki. Istotne cechy regułowej reprezentacji wiedzy to:

– łatwe wykorzystanie takiej reprezentacji w procesach wnioskowania,

– możliwość wykorzystania reguł do różnych celów (np. do wnioskowania, nauczania oraz w procesie objaśnień drogą wnioskowania),
– modularność - każda reguła w bazie wiedzy reprezentuje mały spójny fragment bazy wiedzy,
– przejrzystość bazy wiedzy - reguły można łatwo zapisać wykorzystując język naturalny i łatwo zinterpretować,
– możliwość przyrostowego tworzenia reguł (uzupełniania bazy wiedzy) [8].

1.1. Budowa reguły

Generowanie reguł dotyczy zbioru pojęć K, przy czym pojęcie (ang. concept) K oznacza klasę, zbiór obiektów posiadających pewne wspólne własności. W przypadku modelu klasyfikacyjnego zbiór obiektów U podzielony jest na j podzbiorów Ek1..Ekj, gdzie zbiory spełniają warunek:

$$\forall i \neq j \quad E_{ki} \cap E_{kj} = \varnothing \tag{1}$$

Oraz

$$\bigcup^{r}_{j=1} E_{kj} = U \tag{2}$$

Podział zbioru U dla danego pojęcia K_j oznaczany będzie jako:

$$U = E^{+}_{kj} \cup E^{-}_{kj} \tag{3}$$

I oznaczać będzie podział na dwa podzbiory – podzbiór przykładów pozytywnych:

$$E^{+}_{kj} = E_{kj} \tag{4}$$

oraz podzbiór przykładów negatywnych (tj. pozostałych przykładów):

$$E^{-}_{kj} = U \setminus E_{kj} \tag{5}$$

Reguła decyzyjna **r** definiująca pojęcie K_j zdefiniowana będzie jako:

$$Jeżeli \quad P \quad to \quad Q \tag{6}$$

lub w innym zapisie P→Q, gdzie P jest częścią warunkową reguły (przesłanką) a Q częścią decyzyjną reguły (konkluzją).

Przesłanka P jest koniunkcją (bywa nazywana również kompleksem) warunków elementarnych w i jest reprezentowana w postaci:

$$P = w_1 \wedge w_2 \wedge \ldots \wedge w_k \tag{7}$$

gdzie k jest liczbą wykorzystanych warunków. Liczba k nazywana jest długością reguły i oznaczana Dl(P).

Warunek elementarny w_i (nazywany także selektorem) reguły r jest definiowany jako:

$$(f(a_i, x) \propto term(a_i)) \tag{8}$$

Gdzie $f(a_i,x)$ oznacza wartość atrybutu a_i dla obiektu x, $\propto$ oznacza operator relacji, należącej do zbioru $\{=, \neq, <, \leq, >, \geq, \in\}$, a term oznacza tzw. term elementarny, który jest stałą z dziedziny atrybutu a_i lub może być uogólniony do podzbioru w przypadku operatora $\in$.

Konkluzja Q oznacza, że obiekt x spełniający P należy do K_j. Pokryciem koniunkcji warunków elementarnych P, oznaczanym przez [P], jest zbiór obiektów z tablicy decyzyjnej DT (gdzie DT=(U, A∪{d}). A oznacza atrybuty warunkowe, zaś d reprezentuje atrybut decyzyjny – etykietę klasy), spełniający logiczne warunki reprezentowane przez P.

Pokrycie P można podzielić na część pozytywną:

$$[P]^{+}_{kj} = [P] \cap E^{+}_{kj} \tag{9}$$

oraz cześć negatywną:

$$[P]^{-}_{kj} = [P] \cap E^{-}_{kj} \tag{10}$$

Reguła r jest dyskryminująca, tzn. odróżnia przykłady należące do K_j od przykładów negatywnych wtedy i tylko wtedy, gdy jej przesłanka spełnia warunek niesprzeczności:

$$[P]^{-}_{kj} = \varnothing \tag{11}$$

Wymagane jest również, aby:

$$[P]^{+}_{kj} \neq \varnothing \tag{12}$$

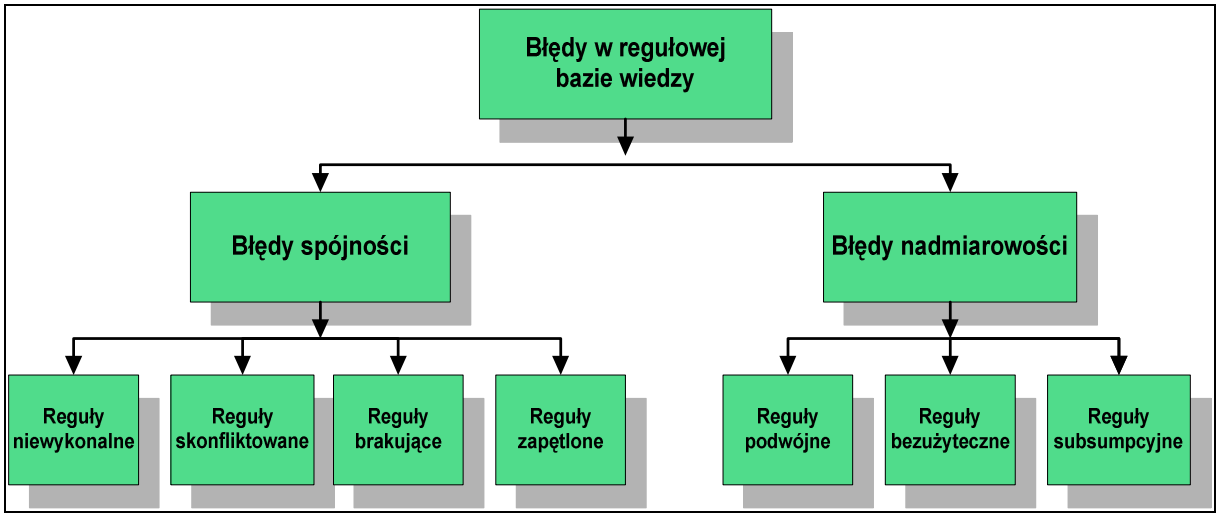
Dyskryminująca **reguła r jest minimalna**, gdy usunięcie jakiegokolwiek warunku elementarnego **w** z jej przesłanki P, prowadzi do sytuacji, w której $[P]^{-}_{kj} \neq \varnothing$.

Zbiór reguł R jest kompletnym opisem klasy K_j, jeżeli każdy przykład $x \in E^{+}_{kj}$ spełnia część warunkową i decyzyjną przynajmniej jednej reguły $r \in R$. [15]

1.2. Błędy w regułowej bazie wiedzy

Uzyskiwany podczas procesu nauczania maszynowego lub z udziałem inżyniera wiedzy zbiór reguł jest często niedoskonały, przy czym wyróżnia się następujące błędy w zbiorze reguł:

- Reguły niewykonalne – zawarte w nich warunki są niemożliwe do spełnienia
- Reguły skonfliktowane – dają wzajemnie sprzeczne konkluzje w tej samej fazie wnioskowania
- Reguły brakujące – w bazie brakuje reguł niezbędnych do wyprowadzenia wniosków finalnych
- Reguły zapętlone – konkluzje z danej reguły wykorzystywane są jako przesłanki w dalszym wnioskowaniu
- Reguły podwójne – powtarzające się reguły o identycznej strukturze
- Reguły bezużyteczne – ich konkluzje nie są wykorzystywane w dalszym wnioskowaniu
- Reguły subsumcyjne – taka para reguł, która prowadzi do tych samych wniosków, przy czym jedna z nich wymaga spełnienia dodatkowych przesłanek



Rys. 1. Rodzaje błędów w regułowej bazie wiedzy.
Źródło: [11]

Po zakończeniu automatycznej budowy bazy wiedzy regułowej powinna zostać przeprowadzona jej weryfikacja, w celu wyszukania ewentualnych błędów. Jednym z dostępnych programów jest np. System Wspomagania Inżynierii Wiedzy CAKE, który umożliwia wykrywanie błędów nadmiarowości, sprzeczności oraz niekompletności. [11].

2. METODY INDUKCJI REGUŁ

Jak wcześniej wspomniano, strategie automatycznego generowania reguł podzielić można na dwie zasadnicze kategorie:

– metody generujące reguły bezpośrednio z danych,
– metody, które opisują wiedzę zgromadzona przez modele lub symulatory, czyli metody które generują reguły na podstawie nauczonych modeli najczęściej klasyfikacyjnych.

Algorytmy oparte o nauczone modele (np. do klasyfikacji) właściwie opisują sposób podejmowania przez nie decyzji. Rozwiązanie takie ma zalety i wady, przy czym do zalet zaliczyć można uniezależnienie od zbioru treningowego, jako, że mając do dyspozycji gotowy model (symulator) można na bieżąco generować nowe przypadki i badać i udoskonalać działanie systemu ekspertowego. Przy czym sam model obiektu może bazować na różnych metodach, od matematycznych (np. modele statystyczne) po metody sztucznej inteligencji (np. sieci neuronowe)

Wśród wad, które mogą zdecydować o niepowodzeniu trzeba zauważyć możliwość nałożenia się na siebie błędu modelu i błędu metody uczenia reguł. Obecnie istnieje wiele algorytmów uzyskiwania reguł np. z sieci neuronowych. Różnią się one między sobą m.in. rodzajem generowanych reguł, ich jakością, liczbą, złożonością samego algorytmu czy sposobem analizy sieci neuronowej. Przykładem mogą być takie systemy jak TREPAN generujący reguły na podstawie odpowiedzi sieci neuronowej na przedstawione jej próbki [2], RuleNet – sieć reguł uczona na podstawie algorytmu opracowanego przez Jacobsa, buduje reguły na podstawie siły powiązań pomiędzy neuronami [więcej w 14]

Niezależnie od wybranej strategii generowania reguł metody indukcji można podzielić na dwie główne kategorie: drzewa decyzyjne i algorytmy pokrywania sekwencyjnego.

Drzewa decyzyjne

Metody drzew decyzyjnych pozwalają na uzyskanie logicznych grafów podejmowania decyzji, które to z kolei łatwo jest przekształcić na zbiór reguł. Są popularne w pakietach komercyjnych takich jak SAS Entreprise, Statistica Data Miner, GhostMiner.[6]

Drzewa klasyfikacyjne pozwalają na jednoznaczne wyrażenie zależności hierarchicznych występujących w procesie podejmowania decyzji, a opis jest czytelny, łatwy w percepcji i weryfikacji. Drzewo składa się z wierzchołków zwanych węzłami oraz krawędzi nazywanych gałęziami. Jeżeli z węzła wychodzą gałęzie skierowane do innych wierzchołków, to taki węzeł nazywany jest rodzicem węzłów, a węzły, do których biegną krawędzie nazywane są dziećmi węzła-rodzica. Węzeł końcowy nazywany jest liściem.

Konstrukcja drzewa oparta jest na indukcji zstępującej (ang. Top-Down Induction of Decision Trees – TDIDT). Przy założeniu, że C – oznacza zbiór uczący, Xl – atrybuty opisujące obiekty ze zbioru uczącego (l=1..m) drzewo generowane jest przez rekurencyjny algorytm o postaci:

```
GenerateTree(C)
begin
If wszystkie przykłady z C należą do jednej klasy then
{
utwórz liść oznaczony nazwą tej klasy
return TREE
        }
        Else
Wybierz jeden atrybut X z wartościami x1..xr i utwórz węzeł X
        Podziel zbiór C na podzbiory C1, C2,..,Cr
        For i=1 to r GenerateTree (Ci)
end
```

Obecnie istnieje wiele metod generowania drzew w wielu odmianach. Różnice pomiędzy algorytmami generowania drzew dotyczą kryterium wyboru atrybutu oraz kryterium stopu. Kryterium stopu może być, jak w przykładzie powyżej, sytuacja gdy wszystkie analizowane przykłady należą już do jednej klasy lub sytuacja gdy zbiór atrybutów ulegnie wy-czerpaniu. [9]

Kryterium wyboru atrybutu to najważniejszy element algorytmu, który stanowi o kolejności atrybutów w drzewie i jego strukturze. Na podstawie tego kryterium klasyfikuje się algorytmy indukcji drzew. Wśród algorytmów tych wyróżnia się m.in. takie, jak CART, ID3,C4.5 [więcej w 1].

Algorytmy pokrywania sekwencyjnego

Ten rodzaj algorytmów jest częście spotykany w oprogramowaniu niekomercyjnym, np. RSES (algorytm Ripper), RapidMiner (algorytm LEM), Weka. [6]

Problem pokrycia wszystkich przykładów ze zbioru uczącego minimalnym zbiorem reguł jest problemem NP-zupełnym (dowód znaleźć można w pracy habilitacyjnej [15]. Z tego powodu do szukania optymalnego zbioru reguł stosowane są algorytmy przybliżone, które opierają się na generowaniu kolejnych pokryć (ang. sequential covering). Polega to na nauczeniu klasyfikatora pojedynczej reguły, a następnie usuwaniu przykładów, które ona pokrywa. Cały proces powtarza się dla pozostałego zbioru obiektów. W efekcie powinien powstać zbiór reguł pokrywający wszystkie przypadki. Poniżej przedstawiony jest za uogólniony algorytm pokrycia:

```
  Sequential covering (Kj klasa, A atrybuty, E przykłady,  próg
akcepta-cji)
  begin
     R:= //zbiór poszukiwanych reguł
     r:=learn-on-rule(Kj klasa, A atrybuty, E zbiór przykładów)
     Jeżeli Evaluate(r,E)>   to
          begin
               R:=R r
               E:=E\[R]  //  usuń  pozytywne  przykłady  pokryte
przez R
               r:=learn-on-rule(Kj  klasa,  A  atrybuty,  E  zbiór
przykła-dów)
          end
     return R
  end
```

Różne algorytmy generowania reguł metodą pokrycia różnią się funkcją learn-on-rule, która może być różnie realizowana oraz funkcją szacowania jakości zbioru reguł - evaluate. przykładów z klasy Kj nie pokrytych przez żadną z wcześniejszych reguł albo do wcześniejszego zakończenia poszukiwań spowodowanych progiem. Cały algorytm powtarza się iteracyjne dla każdej klasy Kj lub do pełnej klasyfikacji. Przedstawiony algorytm jest reprezentantem algorytmów heurystycznych. [7]

Podobnie rzecz ma się podczas wyszukiwania kandydatów na części warunkowe reguł. W zależności od algorytmu stosuje się różne kryteria oceny. Do najczęściej spotykanych należą:

- Maksymalizacja liczby przykładów pozytywnych pokrywanych przez koniunkcję P (tzn. max),
- Maksymalizacja stosunków pozytywnych przykładów pokrywanych przez regułę w odniesieniu do wszystkich przykładów (tzn. max),
- Minimalizację ilości użytych warunków elementarnych (tzn min Dl(P)),
- Inne, np. miara entropii P lub m-estymata prawdopodobieństwa (CESTNIK, 1990).

Podczas konstrukcji przesłanek można korzystać z powszechnie wykorzystywanej w nauczaniu maszynowym (m.in. podczas konstrukcji drzew) zasady minimalnej długości zapisu (ang. Minima Description Length MDL). Wśród wszystkich hipotetycznych możliwości koniunkcji warunków dla danego zbioru przykładów należy wybrać ten, który minimalizuje łączna długość zapisu.

Przykładowymi algorytmami pokryciowymi są algorytm AQ, zaproponowany przez Michalskiego w 1969 omówiony szczegółowo w doktoracie [K. Grąbczewskiego (2003)] Obecnie trwają już prace nad wersją 18. Algorytm posługuje się takimi pojęciami jak: selektory , kompleksy oraz gwiazdy (czyli zbiory kompleksów).

Innym popularnym algorytmem jest PART. Działa podobnie do algorytmu AQ (zasada ang. seperate-and-conguer) generuje regułę pokrywającą grupę obiektów ze zbioru treningowego (część seperate), a następnie generuje kolejną reguły, która pokrywają również część przykładów wcześniej wykorzystywanych (część concquer). Buduje etapami drzewo podobnie jak algorytm C4.5, a ścieżki do najlepszych liści przekształca na reguły[4]

3. OCENA JAKOŚCI REGUŁ

Indukcja reguł klasyfikacyjnych może być prowadzona dla dwóch celów: predykcji klasyfikacji nowych przypadków albo ich opisu. Celem pierwszego podejścia jest zbudowanie zbioru reguł, dzięki któremu możliwe będzie przede wszystkim sprawne klasyfikowanie nowych obiektów, natomiast w drugim podejściu celem jest odkrywanie reguł reprezentujących pewne regularności (lub anomalie), które są interesujące, użyteczne i zrozumiałe. W perspektywie klasyfikacyjnej zbiór reguł jest oceniany pod kątem poprawności klasyfikacji bądź też wrażliwości poszczególnych klas decyzyjnych. W perspektywie opisowej każda reguła oceniania jest indywidualnie, przy czym ocena jest trudniejsza. Może być oparta na subiektywnych wymaganiach użytkownika bądź też na bardziej obiektywnych miarach ilościowych. W praktyce do oceny zbioru reguł stosowane są metody wielokryterialne, agregujące kilka z prostych miar jakości reguł.

Miary ilościowe wiążą regułę ze zbiorem obiektów, na podstawie których została wygenerowana. Przez U oznaczony będzie zbiór obiektów, a reguła ma postać Jeżeli P to Q. Zbiór obiektów spełniających przesłankę oznaczony będzie jako P , zbiór obiektów nie spełniających przesłanki $\neg P$, zbiór obiektów pozytywnych Q , zaś zbiór obiektów negatywnych $\neg Q$ .Większość ilościowych miar oceny reguł wyprowadza się na podstawie tablicy kontyngencji, która zawiera informację o liczebności poszczególnych zbiorów.

Tab. 3. Tablica kontyngencji dla reguł decyzyjnych

	Q	$\neg Q$	Suma
P	n_{PQ}	$n_{P\neg Q}$	n_P
$\neg P$	$n_{\neg PQ}$	$n_{\neg P\neg Q}$	$n_{\neg P}$
Suma	n_Q	$n_{\neg Q}$	n

Źródło: [15]

n_{PQ} - oznacza ilość obiektów spełniających zarówno P jak Q, $n_{P\neg Q}$ - oznacza ilość obiektów spełniających P i nie spełniających Q, itd.

Miary ilościowe dzielą się na trzy kategorie:

- Miary ogólności
- Miary jednostronnej implikacji
- Miary dwustronnej implikacji.

Przykładowe miary ilościowe:

- Miara ogólności części warunkowej:

$$G(P) = \frac{n_p}{n} \tag{13}$$

- Wsparcie reguły (ang. *support*)

$$G(P \wedge Q) = \frac{n_{PQ}}{n} \tag{14}$$

Miary jednostronnej implikacji określają stopień, w jakim prawdziwość P implikuje prawdziwość Q. Im więcej pokrytych przykładów tym lepiej. Możliwy zakres pokrycia to <0,1>

- Bezwzględne wsparcie (ang. *absolute support*)

$$AS(Q \mid P) = \frac{n_{PQ}}{n_P} \tag{15}$$

W niektórych publikacjach określane jest mianem zaufania. Pokazuje stopień w jakim P implikuje Q i przyjmuje wartości ze zbioru <0,1>. Reguła powinna pokrywać jak najmniej przykładów negatywnych. Tych miar ilościowych jest znacznie więcej. [szczegóły w 15].

Innymi miarami, które łatwo jest wprowadzić są

- Pr**ostota reguły** Dl(P), wyrażona przez jej długość (mierzoną ilością warunków elementarnych)
- **Wielkość zbioru reguł** N(R), mierzona całkowitą liczba reguł

Miary złożone powstają poprzez agregacje miar częściowych. Sposób agregacji jest przy tym dowolny. Oto propozycja miary agregacyjnej do oceny zbioru reguł – Kryterium Jakości zbioru reguł:

$$KJ = \frac{(\bar{G}(P) + \bar{A}S(Q \mid P)) \cdot \sigma}{\bar{D}l(P) \cdot N(R)} \tag{16}$$

Jest to miara, którą można zastosować do oceny całego wygenerowanego zbioru reguł. Opiera się na wartościach średnich dla całego zbioru. Wybór poszczególnych miar podyktowany był ich charakterem. Wskazane jest aby reguła pokrywała jak najwięcej przykładów uczących, stąd miara G(P). Pożądane jest również aby wśród pokrywanych przykładów było jak najmniej przykładów negatywnych, stąd miara AS(Q|P). Im reguła jest krótsza tym łatwiejsza do zrozumienia, podobnie jest w przypadku wielkości zbioru reguł. Parametr σ jest dobierany w zależności od wielkości zbioru uczącego i wybranej metody odkrywania wiedzy. Jego zadaniem jest równoważenie wielkości mianownika (który dla $\sigma = 1$ mieści się w zakresie <0,2>, a wielkością mianownika, którego zakres mieści się w przedziale . Z uwagi na te uwarunkowania łatwo określić zakres zmienności miary KJ na $(0; 2 \cdot \sigma >$. Skrajnym przypadkiem będzie zbiór reguł oparty na jednej regule, której przesłanka składać się będzie z 1 warunku, wygenerowana na jednym przykładzie uczącym. Im reguł będzie więcej, im bardziej będą złożone, im mniej wiarygodne tym bardziej miara KJ zbliżać się będzie do 0.[12 i 13]

4. PRAKTYCZNE ZASTOSOWANIA REGUŁOWYCH BAZ WIEDZY

W praktycznych przeznaczeniach uzyskane zbiory reguł podlegają dodatkowemu przetworzeniu zwłaszcza w zastosowaniach klasyfikacyjnych. Algorytmy indukcji reguł do klasyfikacji przeszukujące całą przestrzeń mogą prowadzić do uzyskiwania zbyt skomplikowanych zbiorów reguł. W perspektywie opisu „procesu klasyfikacji" powinno się poszukiwać ogólnych, wiarygodnych i względnie prostych zbiorów reguł. Problem w tym, że zbyt wysoki poziom wsparcia minimalnego prowadzi do ograniczania zbioru do reguł

trywialnych i oczywistych. Z kolei obniżenie wsparcia powoduje gwałtowny wzrost wielkości zbioru reguł.

Aby ograniczyć wielkość zbioru stosuje się techniki ograniczania, dotyczące ilości reguł w zbiorze oraz ich długości. Do technik tych zalicza się filtrację oraz ograniczanie długości przesłanek. Filtracja polega na podzieleniu zbioru reguł zgodnie z przyjętym kryterium, np. zaufaniem do reguły czy obecnością konkretnego atrybutu w przesłance. Wymaga wcześniej wygenerowania pełnego zbioru reguł lub też może odbywać się podczas samego procesu generowania reguł (np. konstrukcji drzewa decyzyjnego). Takie ograniczenie zbioru reguł pozwala na uzyskanie prostszych zbiorów reguł, często ograniczenie liczby atrybutów opisowych w przesłankach, a w efekcie uproszczenie opisu klasyfikowanych obiektów jak i samego procesu klasyfikacji. Jednocześnie należy być świadomym, że taka redukcja to zawsze jest rozwiązanie, które obniża skuteczność klasyfikacji. Kompromis polega na znalezieniu poziomu błędu dopuszczalnego dla przewidywanego zastosowania bazy wiedzy.

Przykład pierwszy

Pierwszym przykładem zastosowania regułowej bazy wiedzy jest baza do klasyfikacji polskich spółek giełdowych przeprowadzona wg 6-cio stopniowej skali ratingu międzynarodowego. Badanie zostało wykonane na bazie sprawozdań finansowych za lata 1998-2003, atrybutami opisowymi były powszechnie znane wskaźniki finansowe (w liczbie 30). Baza indukowano na podstawie rzeczywistych danych finansowych (bezpośrednia obserwacja eksperta). Przebadano różne algorytmy (drzewa m.in. ID3, CART, algorytmy pokrywania sekwencyjnego min. PART, zbiory przybliżone)indukcji baz regułowych pod kątem dokładności klasyfikacji oraz pod kątem jakości uzyskanego zbioru reguł mierzonego miarą wg wzoru (16). Liczba reguł ograniczania była minimalnym poziomem wsparcia reguły: od 10 do 100 przypadków oraz minimalnym poziomem zaufania do reguł: 0,25.

Przeprowadzone eksperymenty dowiodły, że najwyższą skuteczność klasyfikacji osiągały algorytmy generujące reguły o wsparciu minimalnym na poziomie 10 instancji- odpowiednio ID3 54,5%, CART 52,10%, C4.5 65,8%, PART 54,2% i uzyskując miary jakości (w nawiasie podano liczbę reguł) ID3 0,31 (58), CART 36,42 (10), C4.5 15,31 (15), PART 0,37 (61).

Jednakże dla minimalnego poziomu wsparcia równemu 50 instancji wyniki algorytmów były następujące: skuteczność klasyfikacji: Id3 46%, CART 52,5%, PART 55,4% i odpowiednio jakość zbioru reguł ID3 18,27 (14), CART 17,42 (14), PART 14,84 (13).

W tym drugim przypadku dla części algorytmów drzew takich jak ID3 wzrosła jakość zbioru reguł ale kosztem istotnego spadku dokładności klasyfikacji, a dla drugiej części jakość zbioru reguł nawet spadła. Natomiast dla algorytmów pokryciowych (PART) uzyskano istotny wzrost kryterium jakości reguł, a towarzyszy temu nawet niewielki wzrost skuteczności klasyfikacji. [badania własne, więcej w 16 i 17] Doświadczenie to dowodzi, że nie ma jednej najskuteczniejszej metody indukcji bazy wiedzy, a także że niejednokrotnie dążenie do uproszczenia zbioru reguł może przyczyniać się do zwiększania skuteczności działania systemu ekspertowego.

Przykład drugi

Drugim przykładem będzie badanie przeprowadzone przez D.A. Pearce. Indukował on dwie bazy reguł dla systemu ekspertowego do wspomagania diagnostyki układu zasilania satelity na podstawie wartości pomiarów urządzeń telemetrycznych. Z uwagi na wysoki koszt obserwacji bezpośredniej badanie zachowania satelity zostało wykonane na symulatorze czasu rzeczywistego urządzenia, wykorzystywanym do trenowania członków obsługi satelity. Jedna z baz reguł została wygenerowana manualnie na podstawie współpracy z rzeczywistym ekspertem. Po wielu symulacjach wszystkich wariantów awarii powstał w ten sposób klasyczny system ekspertowy do identyfikacji awarii. Druga baza reguł została wygenerowana metodą uczenia maszynowego (algorytm AQ w odmianie AQR) na podstawie

ponad 700 wywołanych przypadków awarii. W efekcie powstał zbiór 75 reguł opierających się średnio na 3 przesłankach (symptomach), gdzie każda z awarii początkowo opisana była 30 atrybutami. Obie bazy zostały porównane na podstawie wygenerowanych na symulatorze objawów 14 awarii z możliwie szerokiego zakresu. Baza indukowana automatycznie uzyskała 100% skuteczność w diagnozowaniu awarii, natomiast klasyczny system ekspertowy prawidłowo rozpoznał 10 przypadków (skuteczność 72%). [10] Doświadczenie to dowodzi, że posługiwanie się metodami automatycznej indukcji reguł może przynieść lepsze rezultaty w skuteczności działania systemów ekspertowych niż tradycyjne metody pozyskania baz wiedzy. A także przekonuje o możliwości wykorzystywania modeli (symulatorów) do zbierania „doświadczenia" eksperckiego.

PODSUMOWANIE

Przedstawione w artykule zagadnienia związane z pozyskiwaniem regułowych baz wiedzy pokazują, że jest to złożony proces, wymagający uważnego zbadania wszystkich możliwości indukcji reguł, które są dostępne w konkretnym przypadku. Problemem może być zarówno brak ekspertów dla danej dziedziny jak i duża złożoność badanego systemu, która zwykle prowadzi do błędów logicznych reguł eksperckich lub braków w obszarach pokrycia wiedzy. W uzasadnionych przypadkach korzystniejszym rozwiązaniem może być wykorzystanie modelu lub symulatora obserwowanego systemu, paradoksalnie może doprowadzić do uzyskania w efekcie skuteczniejszych systemów ekspertowych.

BIBLIOGRAFIA

1. Breiman L. Friedman R. Ohlsen A, Stone C. *Classification and regression trees*, Blemont, California Wadsworth International Group 1984
2. Craven M.W., *Exctracting comprehensible models from trained neural networsk*, Praca doktorska, Madison, University of Wisconsin,1996 http://www.biostat.wisc.edu/~craven/papers/thesis.pdf październik 2013
3. Cestnik, B. Estimating Probabilities: a crucial task in machine learning, Materiały konferencyji ECAO, Sztokholm, 1990
4. Eibe F. ,. Witten I. H. (1998): Generating Accurate Rule Sets Without Global Optimization., Machine Learning: Proceedings of the Fifteenth International Conference, styczeń 2006 www.cs.waikato.ac.nz/~eibe/pubs/ML98-57.ps. styczeń2009
5. Grąbczewski K, *Zastosowanie kryterium separowalności do generowania reguł klasyfikacji na podstawie baz danych*, Praca doktorska, Toruń, Uniwersytet Mikołaja Kopernika, 2003
6. W.Malara, W.Sikora, Ł. Wróbel, *Program do indukcji i oceny reguł klasyfikacyjnych, zintegrowany z pakietem R*, Studia Informatica, vol. 34 2013 http://www.znsi.aei.polsl.pl/materialy/SI112/SI112_26.pdf październik 2013
7. Mitchell T, *Machine Learning*, Boston, MacGraw Hill 1997
8. Mrózek A., Płonka L., *Analiza danych metodą zbiorów przybliżonych. Zastosowanie w ekonomii, medycynie i sterowaniu*. Warszawa, Akademicka Oficyna Wydawnicza PLJ, 1999
9. Nycz M. *Generowanie wiedzy dla przedsiębiorstwa*, Wrocław, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, 2004
10. D.A. Perace, *The induction of Fault Diagnosis Systems for Qualitive Models*, The Turing Institute 36 North Hanover St, Glasgow, Scotland http://www.aaai.org/Papers/AAAI/1988/AAAI88-063.pdf, październik 2013

11. Radosiński E., *Systemy informatyczne w dynamicznej analizie decyzyjnej*, Warszawa-Wrocław, PWN 2001
12. Rozenberg L., Trojczak-Golonka P., *Metoda odkrywania wiedzy w bazach danych a jej jakość. Studium przypadku na przykładzie danych finansowych spółek polskich*, Materiały konferencji Problemy Społeczeństwa Informacyjnego, Międzyzdroje, Uniwersytet Szczeciński, 2007
13. Rozenberg L., Trojczak-Golonka P.*Impact of attributes selection on the quality of the enterprise classification on the example of the relief type algorithm*, Materiały Ogólnopolskiej Konferencji Sejmik Młodego Informatyka, Świnoujście, 2007
14. Sienkiewicz W., Sieci neuronowe w tworzeniu reguł w systemach ekspertowych, wrzesień 2006
http://aragorn.pb.bialystok.pl/~radev/ai/sosn/sienkiewicz.htm luty 2013
15. Stefanowski J., *Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy*, Rozprawa habilitacyjna, Poznań, Politechnika Poznańska, 2001
luty 2013 www-idss.cs.put.poznan.pl/~stefan/pub/habjs.pdf 2
16. Nadolna B., Trojczak-Golonka P., *„The meaning of attributes ranking in the multi-state classification of dynamic objects”*, "Polish Journal of Enviromental Studie", vol. 17, No. 3B, 2008
17. Trojczak-Golonka P., *„Metoda syntezowania wzorców diagnostycznych z optymalizacją liczby atrybutów”*, rozprawa doktorska, Zachodniopomorski Uniwersytet Technologiczny, Szczecin 2007

PRINCIPLES OF THE AUTOMATIC INDUCTION RULE-BASED KNOWLEDGE BASES FOR EXPERTS SYSTEMS

Abstract

In the paper the advantages and disadvantages of rules-based representation of the knowledge base and the basic class methods of machine induction were presented. There are shown the problems related to such process, the application examples and using areas rule-based knowledge bases

Autorzy*:*
dr inż. **Patrycja Trojczak-Golonka** – Akademia Marynarki Wojennej – Wydział Nawigacji i Uzbrojenia Okrętowego.
dr inż. **Stanisław Milewski** – Akademia Marynarki Wojennej – Wydział Nawigacji i Uzbrojenia Okrętowego.