

## UCZENIE PARAMETRÓW SIECI BAYESOWSKICH Z DANYCH Z WYKORZYSTANIEM BRAMEK NOISY-OR

**Agnieszka Oniśko\***, **Marek J. Druzdziel\*\***, **Hanna Wasyluk\*\*\***

\* Wydział Informatyki, Politechnika Białostocka, Wiejska 45A, Białystok, 15-351

\*\* Decision Systems Laboratory, School of Information Sciences, and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA

\*\*\* Centrum Medycznego Kształcenia Podyplomowego, Marymoncka 99, Warszawa, Instytut Biocybernetyki i Inżynierii Biomedycznej PAN

Zbiory danych mogą istotnie ułatwić parametryzację sieci bayesowskich. Niestety w przypadku, gdy liczność zbioru jest niewielka, niektóre z kombinacji wartości w tabelach, reprezentujących warunkowe rozkłady prawdopodobieństwa, są reprezentowane przez kilka lub też zerowa liczbę rekordów. W takiej sytuacji dane nie pozwalają na wyznaczenie wiarygodnych parametrów. W poniższym artykule, proponujemy metodę, która wykorzystuje bramki Noisy-OR w uczeniu parametrów sieci bayesowskiej z danych. Zaproponowana metoda została przetestowana przy użyciu modelu HEPAR II diagnozującego choroby wątroby, którego parametry zostały wyznaczone na podstawie medycznego zbioru danych. Jakość diagnostyczna modelu, którego parametry zostały „wygładzone” przy wykorzystaniu parametrów Noisy-OR była lepsza o 6,7%.

### 1. Wprowadzenie

Sieci bayesowskie, (Pearl, 1988), nazywane również probabilistycznymi modelami graficznymi, sieciami przekonań lub sieciami przyczynowo-skutkowymi, stały się na przełomie ostatniej dekady popularnym narzędziem do reprezentacji wiedzy w warunkach niepewności, (Henrion i in., 1991). Sieć bayesowska jest acyklicznym grafem skierowanym i składa się z części jakościowej, która stanowi zbiór zmiennych – węzłów grafu wraz z probabilistycznymi zależnościami pomiędzy nimi oraz części ilościowej sieci, reprezentującej rozkład prawdopodobieństwa łącznego dla tych zmiennych. Z punktu widzenia inżynierii wiedzy, sieć bayesowska może odzwierciedlać strukturę przyczynowo-skutkową, która pozwala na pełniejsze zrozumienie modelowanego problemu zarówno przez ekspertów jak i użytkowników systemu.

Głównym elementem w budowaniu sieci bayesowskiej jest określenie jej struktury oraz parametryzacja. Kompletna tabela reprezentująca rozkład prawdopodobieństwa warunkowego dla zmiennej binarnej z  $n$  binarnymi poprzednikami w sieci bayesowskiej wymaga  $2^n$  niezależnych parametrów. W przypadku znaczących wartości  $n$ , określenie  $2^n$  parametrów przez eksperta może okazać się trudne, a czasami nawet niemożliwe ze względu na ograniczony czas, jakim dysponują eksperci. Niewątpliwą zaletą sieci bayesowskich jest fakt, że pozwalają one na łączenie wiedzy eksperta z danymi. Jeżeli odpowiednia liczba danych jest dostępna, sieci bayesowskie, zarówno ich struktura, jak i parametry, mogą być nauczone z danych, (Cooper i Herskovits, 1992), (Pearl i Verma, 1991), (Spirtes i in., 1993). Jakkolwiek, w przypadku zbiorów danych z niewielką liczbą rekordów, jakość tych modeli jest niska.

Głównym przedmiotem tego artykułu jest uczenie parametrów warunkowych rozkładów prawdopodobieństw (WRP) w sieciach bayesowskich z niewielkich zbiorów danych dla istniejącej struktury modelu. Uczenie parametrów sieci bayesowskiej sprowadza się głównie do zliczania liczby rekordów dla różnych warunków (kombinacji stanów parametryzowanego węzła i jego rodziców). Rozkłady a priori (dla węzłów, które nie posiadają bezpośrednich poprzedników) mogą być w miarę wiarygodnie uczone z danych, natomiast wyznaczanie WRP jest trudniejsze. W przypadku zbiorów z niewielką liczbą rekordów, niejednokrotnie dana kombinacja wartości w tabeli WRP jest reprezentowana przez zaledwie kilka rekordów (bądź też nawet przez zerową liczbę rekordów), które nie pozwalają na wyznaczenie wiarygodnych parametrów. W takich sytuacjach często stosuje się rozkłady jednostajne. W poniższym artykule proponujemy proces wygładzania WRP, które są uczone z danych, poprzez łączenie danych ze strukturalnymi i numerycznymi informacjami pochodzącymi od eksperta. Informacje te dotyczyły wiedzy eksperta, który wskazał te zmienne, dla których rozkład WRP może być przybliżony przy użyciu bramki Noisy-OR, Henrion (1989), Pearl (1988). Dla węzłów tych wyznaczaliśmy dwa zbiory parametrów Noisy-OR: (1) parametry pozyskane od eksperta i (2) parametry nauczone z danych. Równolegle dla każdego węzła wyznaczaliśmy WRP bezpośrednio z danych. W sytuacji, gdy znaleźliśmy odpowiednią liczbę rekordów dla danej kombinacji wartości w tabeli warunkowego rozkładu prawdopodobieństwa, wyznaczaliśmy parametry na ich podstawie. Z kolei, gdy liczba rekordów była niewystarczająca, wówczas generowaliśmy WRP na podstawie parametrów Noisy-OR.

Bramki Noisy-OR były stosowane w medycznych modelach sieci bayesowskich (np. Diez i in. (1997), Shwe i in. (1991)), jakkolwiek metoda, którą proponujemy jest nowa. Zaproponowaną metodę przetestowaliśmy dla HEPAR II, modelu sieci bayesowskiej dla diagnozowania chorób wątroby. Model ten składa się z 73 węzłów, jego parametry zostały wyznaczone na podstawie zbioru danych składającego się z 505 przypadków chorobowych. Wyniki przeprowadzonych eksperymentów dowodzą, że zaproponowana metoda prowadzi do polepszenia jakości diagnostycznej modelu.

## 2. Bramki Noisy-OR

Niektóre rozkłady prawdopodobieństw warunkowych mogą być aproksymowane przez interakcyjne modele kanoniczne, które wymagają mniejszej liczby parametrów. Niejednokrotnie rozkłady te mogą wiarygodnie przybliżyć prawdziwy rozkład, jak również istotnie zredukować wysiłek związany z budowaniem modelu. Jednym z rozkładów kanonicznych, stosowanym w sieciach bayesowskich są bramki Noisy-OR, (Diez i Druzdzel, 2002), (Henrion, 1989), (Pearl, 1988). Bramki Noisy-OR są zazwyczaj stosowane w sytuacji związanej z opisem oddziaływań pomiędzy  $n$  przyczynami  $X_1, X_2, \dots, X_n$  oraz ich wspólnym efektem  $Y$ . Przy czym zakłada się, że każda z przyczyn jest w stanie prowadzić do efektu  $Y$ , w sytuacji, gdy pozostałe przyczyny są nieobecne oraz gdy ich zdolność do wywoływania efektu  $Y$  jest niezależna od występowania pozostałych przyczyn. Jednym z najprostszych modeli kanonicznych jest binarna bramka Noisy-OR, (Pearl, 1988), którą można zastosować w sytuacji, gdy istnieje kilka przyczyn  $X_1, X_2, \dots, X_n$  zmiennej reprezentującej efekt  $Y$ . Poza tym, każda z przyczyn jest charakteryzowana przez prawdopodobieństwo  $p_i$ , które mówi, że jest ona w stanie powodować efekt w sytuacji, gdy pozostałe przyczyny są nieobecne. Dodatkowo, zdolność do wywoływania efektu przez daną przyczynę jest niezależna od obecności pozostałych przyczyn. Powyższe założenia pozwalają dokonać specyfikacji rozkładu prawdopodobieństwa warunkowego dla danej zmiennej na podstawie  $n$  parametrów  $p_1, p_2, \dots, p_n$ .  $p_i$  reprezentuje prawdopodobieństwo tego, że efekt  $Y$  wystąpi, jeśli przyczyna  $X_i$  jest obecna i pozostałe przyczyny  $X_j$ , (dla  $j \neq i$ ) są nieobecne.

$$p_i = \Pr(y | \bar{x}_1, \bar{x}_2, \dots, \bar{x}_i, \dots, \bar{x}_{n-1}, \bar{x}_n) \quad (1)$$

Stąd, prawdopodobieństwo  $y$  pod warunkiem  $X_p$  (podzbiór tych zmiennych  $X_i$ , które są obecne) opisane jest równaniem:

$$\Pr(y | X_p) = 1 - \prod_{i: X_i \in X_p} (1 - p_i) \quad (2)$$

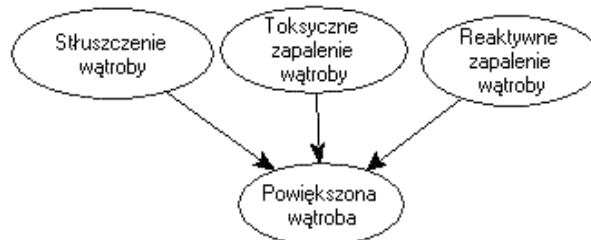
Równanie to pozwala na wyznaczenie kompletnej tabeli prawdopodobieństw warunkowych zmiennej  $Y$  pod warunkiem jej poprzedników  $X_1, X_2, \dots, X_n$ .

(Henrion, 1989) zaproponował poszerzenie binarnej bramki Noisy-OR o sytuację, w której efekt występuje nawet wtedy, jeśli wszystkie modelowane przyczyny są nieobecne. Model ten nazwał bramką Noisy-OR z przeciekiem (*ang. leaky Noisy-OR*). Bramka Noisy-OR z przeciekiem ma szczególne zastosowanie w sytuacjach, kiedy model nie uwzględnia wszystkich możliwych przyczyn efektu  $Y$  (przypadek większości modeli).

Sytuacja taka może być modelowana przez wprowadzenie dodatkowego parametru  $p_0$ , który będzie reprezentować prawdopodobieństwo przecieku (*ang. leak*), tzn. łączny efekt wszystkich niemodelowanych przyczyn zmiennej  $Y$ .

$$p_0 = \Pr(y | \bar{x}_1, \bar{x}_2, \dots, \bar{x}_n) \quad (3)$$

$p_0$  reprezentuje prawdopodobieństwo tego, że efekt  $Y$  wystąpi, pomimo, że wszystkie modelowane przyczyny będą nieobecne.



**Rysunek 1: Przykład bramki Noisy-OR**

Rysunek 1 przedstawia przykład bramki Noisy-OR dla węzła *Powiększona wątroba*. Każdy z rodziców tego węzła, tzn. *Stłuszczenie wątroby*, *Toksyczne zapalenie wątroby* i *Reaktywne zapalenie wątroby*, może niezależnie powodować powiększenie wątroby. Wątroba może również ulec powiększeniu pod wpływem innych niemodelowanych w tym przykładzie czynników. W przypadku bramki Noisy-OR z przeciekiem, prawdopodobieństwo  $p_i$  ( $i \neq 0$ ) nie reprezentuje już prawdopodobieństwa tego, że  $X_i$  prowadzi do wystąpienia  $Y$  pod warunkiem, że pozostałe przyczyny są nieobecne. Wartość ta oznacza prawdopodobieństwo tego, że  $Y$  wystąpi gdy  $X_i$  jest obecne i nie są obecne inne pozostałe modelowane przyczyny  $X_j$  dla  $j \neq i$ .

Niech  $p_i'$  będzie prawdopodobieństwem tego, że  $Y$  wystąpi, jeśli  $X_i$  jest obecne i pozostałe przyczyny zmiennej  $Y$ , włączając niemodelowane przyczyny, są nieobecne.  $p_i'$  reprezentuje prawdopodobieństwo tego, że  $X_i$  powoduje  $Y$ . Niech:

$$1 - p_i' = \frac{1 - p_i}{1 - p_0} \quad (4)$$

Stąd mamy:

$$p_i = p_i' + (1 - p_i')p_0 \quad (5)$$

Prawdopodobieństwo  $Y$  pod warunkiem  $X_p$  (zbiór tych przyczyn  $X_i$ , które są obecne) dla bramki Noisy-OR z przeciekiem wyrażone jest następującym równaniem:

$$\Pr(Y | X_p) = 1 - (1 - p_0) \prod_{i: X_i \in X_p} \frac{1 - p_i}{1 - p_0}$$

Diez (1993) zaproponował alternatywne rozwiązanie pozyskiwania parametrów Noisy-OR z przeciekiem. Jego metoda sprowadzała się do zapytania eksperta o wartość parametrów  $p_i'$  (patrz równanie (4)). Różnica pomiędzy dwiema metodami związana była z parametrem przecieku. Parametry Henrion'a zakładały, że odpowiedź eksperta zawiera jednocześnie prawdopodobieństwo wystąpienia efektu pod wpływem przyczyny oraz niemodelowanego przecieku. Parametry Diez'a  $p_i'$  reprezentują mechanizm, jaki zachodzi między przyczyną i skutkiem w sytuacji, gdy przeciek jest nieobecny. Konwersji pomiędzy parametrami Henrion'a i Diez'a dokonać można na podstawie równania (5).

Rozszerzenie binarnego modelu Noisy-OR do modelu z niebinarnymi zmiennymi zaproponowali Diez (1993), Henrion (1989), i Srinivas (1993). W artykule tym sięgnęliśmy po definicje zaproponowane przez Diez'a i Henrion'a. Po szczegółowy opis zagadnień związanych z rozkładami kanonicznymi odsyłamy czytelników do oryginalnych artykułów oraz do pracy Diez'a i Druzdzeła (2002).

### 3. Model HEPAR II i kliniczna baza danych HEPAR

Wspieranie diagnostyki chorób wątroby było przedmiotem wielu prac, np. (Adlassaning i Horak, 1995), (Bobrowski, 1992), (Lucas i in., 1989, 1994), (Richards i in., 1996). Wykorzystując techniki analizy decyzji zbudowaliśmy model sieci bayesowskiej, HEPAR II, którego zadaniem jest wspomaganie decyzji lekarza w diagnozowaniu chorób wątroby, (Oniśko i in., 2000, 2001). System HEPAR II jest kontynuacją projektu HEPAR, który powstał w Instytucie Biocybernetyki i Inżynierii Biomedycznej PAN we współpracy z lekarzami z Centrum Medycznego Kształcenia Podyplomowego w Warszawie, (Bobrowski, 1992), (Wasyluk, 1995). System został zaprojektowany w celu gromadzenia i przetwarzania danych klinicznych pacjentów z chorobami wątroby. Głównym

zadaniem HEPAR'u było zredukowanie ilości wykonywanych biopsji wątroby, która stanowi inwazyjne i dość kosztowne badanie pacjenta. System HEPAR jest zintegrowany z bazą danych, która powstała w 1990 i jest używana w Klinice Gastroenterologii Instytutu Żywności i Żywienia w Warszawie. Każdy przypadek hepatologiczny w bazie HEPAR jest opisany przez ok. 160 różnych zmiennych medycznych, takich jak symptomy, wyniki badań przedmiotowych i podmiotowych, wyniki testów laboratoryjnych oraz histopatologicznie zweryfikowaną diagnozę.

Jednym z założeń analizowanego zbioru danych był fakt, że każdy z pacjentów był związany tylko z jedną jednostką chorobową. Czyli, że wszystkie jednostki chorobowe wzajemnie się wykluczają. Założenie to sprawiło, że pierwsze wersje modelu HEPAR II odzwierciedlały to założenie, (Oniśko i in., 2000). Struktura modelu HEPAR II została zbudowana w oparciu o literaturę medyczną oraz wywiady z ekspertami w dziedzinie hepatologii. Wersja modelu HEPAR II, której użyliśmy w naszych eksperymentach składała się z 73 węzłów reprezentujących 9 różnych chorób wątroby, 66 zmiennych stanowiących wyniki badań przedmiotowych i podmiotowych oraz testów laboratoryjnych. Szacujemy, że budowanie struktury zajęło w przybliżeniu 60 godzin spędzonych na wywiadach z ekspertami. Parametry numeryczne modelu zostały wyznaczone na podstawie zbioru danych HEPAR.

Wartości brakujące stanowią problem w estymacji parametrów z danych. Aby wyznaczyć parametry naszego modelu z danych, przyjęliśmy następujące założenie. W przypadku zmiennych reprezentujących symptomy, wyniki badań przedmiotowych i podmiotowych, wartości brakujące zostały potraktowane jako nieobecne. Z kolei wyniki testów laboratoryjnych zostały zastąpione wartością reprezentującą normę. Podejście to oparliśmy na pracy Peot i Shachter (1998), w której to zaobserwowano, że wartości brakujące w medycznych zbiorach danych nie są wartościami brakującymi losowo i zazwyczaj związane są z wartościami reprezentującymi normę. Należy zauważyć, że w przypadku wnioskowania w sieci bayesowskiej do zdiagnozowania nowego pacjenta nie musimy znać wszystkich wartości zmiennych opisujących jego stan.

Wnioskowanie w sieci bayesowskiej sprowadza się do wyznaczenia rozkładu prawdopodobieństwa a posteriori pod warunkiem zaobserwowanych wartości zmiennych modelu. Rozkład tego prawdopodobieństwa może być bezpośrednio wykorzystany we wspomagananiu decyzji diagnostycznych.

#### 4. Specyfikacja parametrów Noisy-OR

Dla każdego węzła i jego bezpośrednich poprzedników zweryfikowaliśmy z udziałem eksperta czy rozkład danego węzła może być aproksymowany przez bramkę Noisy-OR. Ekspert zidentyfikował 25 węzłów (spośród 62 węzłów z rodzicami), które mogły być przybliżone przez bramki Noisy-OR. Interakcja może być przybliżona przez bramkę Noisy-OR, jeśli spełnia ona następujące założenia: (1) zarówno zmienna reprezentująca efekt, jak i przyczyny powinny reprezentować obecność jakiejś nieprawidłowości (czyli np. zmienne *wiek* lub *pleć* nie mogłyby występować w bramce Noisy-OR), (2) każdy z rodziców musi reprezentować przyczynę, która może powodować wystąpienie efektu w nieobecności pozostałych przyczyn, (3) nie powinno być istotnych oddziaływań pomiędzy przyczynami. Proponowana nowa metoda uczenia parametrów związana jest z wygładzaniem parametrów, tzn. każdy z węzłów, który został rozpoznany jako bramka Noisy-OR, był przedmiotem następującej operacji: w sytuacji, gdy znaleźliśmy odpowiednią liczbę rekordów dla danej kombinacji

wartości w tabeli WRP, wyznaczaliśmy parametry na ich podstawie. Z kolei, gdy liczba rekordów była niewystarczająca, wówczas generowaliśmy WRP na podstawie wcześniej wyznaczonych parametrów Noisy-OR. Założenie, które przyjęliśmy mówiło, że WRP będzie najlepiej odzwierciedlać właściwy rozkład, z kolei rozkład Noisy-OR będzie lepszy niż rozkład jednostajny w sytuacji, gdy nie znajdziemy odpowiednio dużej liczby rekordów. Poniższe dwie sekcje opisują, w jaki sposób zostały określone parametry Noisy-OR.

#### 4.1 Uczenie parametrów Noisy-OR z danych

Wartości parametrów Noisy-OR dla każdej z 25 bramek Noisy-OR zostały wyznaczone na podstawie przecieku równania (1). Z kolei parametry rwyznaczono na podstawie równania (3). Próbowaliśmy również wyznaczać parametry Noisy-OR przez dopasowywanie rozkładu Noisy-OR do większego fragmentu tabeli prawdopodobieństwa warunkowego, jakkolwiek, najprostsze rozwiązanie przyniosło najlepsze rezultaty.

#### 4.2 Pozyskanie parametrów Noisy-OR od eksperta

Dodatkowo, dla każdej z 25 bramek Noisy-OR pozyskaliśmy parametry Noisy-OR od eksperta. Pozyskanie 189 parametrów zajęło około czterech godzin. Przed pozyskaniem parametrów zadaliśmy ekspertowi dwa rodzaje pytań, które związane były z zaproponowanymi w literaturze formalizmami dla bramek Noisy-OR. Pierwszy rodzaj pytania dotyczył parametrów  $p_i$  (patrz równanie (1)) i dotyczył definicji Henriona (1989). Zadanie pytania dla przykładowej bramki (patrz rysunek 1) sprowadzałoby się do następującej formuły:

Jakie jest prawdopodobieństwo tego, że *Toksyczne zapalenie wątroby* powoduje *Powiększenie wątroby*, jeżeli nie zaobserwowano ani *Stłuszczenia wątroby* ani *Reaktywnego zapalenia wątroby*?

Drugi rodzaj pytań związany był z parametrami  $p_i'$  (patrz równanie (4)) i dotyczył definicji Diez'a (1993). Zadanie pytania dla przykładowej bramki (patrz rysunek 1) sprowadzałoby się do następującej formuły:

Jakie jest prawdopodobieństwo tego, że *Toksyczne zapalenie wątroby* powoduje *Powiększenie wątroby*, jeżeli żadna inna z przyczyn *Powiększenia wątroby* jest nieobecna?

W trakcie pozyskiwania parametrów od eksperta zauważyliśmy, że ekspert preferuje definicję zaproponowaną przez Diez'a.

**Tabela 1: Parametry Noisy-OR pozyskane od eksperta dla zmiennej *Cholesterol całkowity***

Cholesterol/Rodzic	Stłuszczenie	AZW	PZW	PBC	Przeciek
Wysoki	0,02	0,02	0	0	0
Średni	0,6	0,4	0,25	0,1	0,01
Norma	0,38	0,58	0,75	0,9	0,99

Podczas pozyskiwania parametrów od eksperta, zaobserwowano, że ekspert preferuje najpierw określać parametry albo dla stanów reprezentujących normę, albo dla tych, które związane są ze skrajnie nieprawidłowym stanem. Tabela 1 zawiera parametry Noisy-OR, które określił ekspert

(skrótów zastosowane w tabeli: AZW, PZW, i PBC oznaczają odpowiednio: *Aktywne zapalenie wątroby*, *Przetrwałe zapalenie wątroby* oraz *Pierwotną marskość żółciową*). Ekspert zazwyczaj podawał najpierw wstępne wartości, a później modyfikował je dopóty, dopóki nie osiągnął satysfakcjonujących go wartości. Ekspert preferował określanie parametrów Noisy-OR w obrębie danego stanu, np. dla wartości „Norma” określał najpierw prawdopodobieństwa dla każdej z przyczyn. Zauważyliśmy, że ekspert ma pewne trudności w określaniu wartości przecieku, tzn. ekspert często podawał wartości dla populacji szpitalnej, zamiast wartości reprezentujące populację generalną.



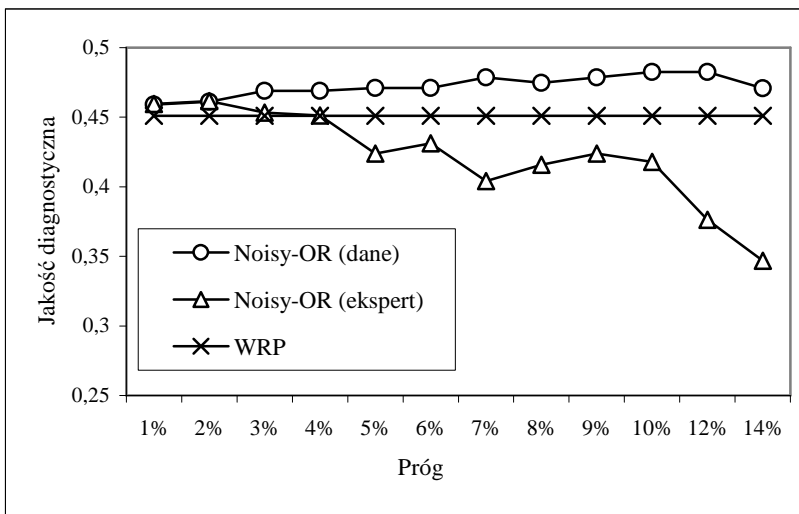
**Rysunek 2: Procent zastąpionych wartości w tabelach WRP jako funkcja progu**

### 5. Porównanie jakości diagnostycznej modeli

Przeprowadziliśmy szereg eksperymentów związanych z testowaniem jakości diagnostycznej różnych wersji modelu HEPAR II. Parametry każdej z wersji uczone były na podstawie tego samego zbioru danych. Zbiór danych składał się z 505 rekordów, reprezentujących przypadki chorobowe. Każdy z pacjentów był sklasyfikowany do jednej z dziewięciu jednostek chorobowych wątroby. Dla każdej wersji modelu zastosowaliśmy tę samą miarę jakości, tzn. diagnostyczna jakość była wyznaczana na podstawie metody *leave-one-out* (Moore i Lee, 1994). Przez jakość diagnostyczną rozumiemy stosunek pacjentów, którzy zostali poprawnie zdiagnozowani, do wszystkich przypadków chorobowych w zbiorze danych. Diagnoza dla każdego pacjenta była wyznaczona poprzez wprowadzenie do modelu wartości zaobserwowanych cech, tzn. zbiór 66 możliwych cech reprezentujących symptomy, wyniki badań przedmiotowych i podmiotowych, oraz testy laboratoryjne. Dane te nie zawierały wyników biopsji. Obserwowane były tylko te cechy, dla których wartości były określone. W eksperymentach wzięliśmy pod uwagę trzy wersje modelu HEPAR II: (1) model, którego parametry numeryczne zostały wyznaczone bezpośrednio na podstawie zbioru danych i dwa modele, których parametry zostały wygładzone przez: (2) parametry Noisy-OR wyznaczone z danych, oraz (3) parametry Noisy-OR pozyskane od eksperta.

Proces wygładzania parametrów związany był z zastępowaniem tych elementów tabeli WRP, dla których nie znaleźliśmy odpowiedniej liczby rekordów, tzn. jeśli liczba rekordów była niższa od wartości progu (próg określiliśmy jako procent wszystkich rekordów w zbiorze danych, np. próg

równy 10% odpowiada w przybliżeniu 50 rekordom). Rysunek 2 prezentuje zależność między progiem a procentem wszystkich wartości w tabeli WRP, które zostały zastąpione przez rozkład Noisy-OR. Procent zastąpionych prawdopodobieństw był wprost proporcjonalny do wartości progu.



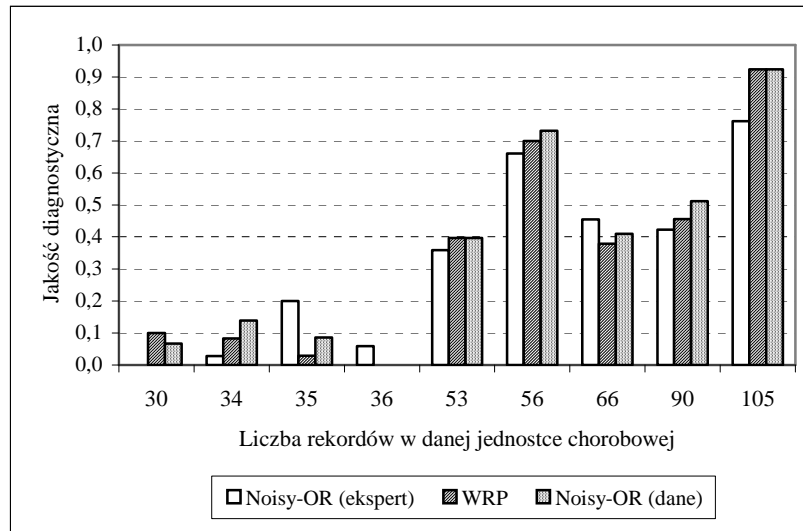
**Rysunek 3: Jakość diagnostyczna modeli jako funkcja progu**

Rysunek 3 prezentuje rezultaty dla trzech modeli. Przedstawia on jakość diagnostyczną modeli jako funkcję progu. Najwyższą jakość diagnostyczną osiągnął model, którego parametry zostały wygładzone przez parametry Noisy-OR wyznaczone z danych. Najwyższą jakość modeli zaobserwowaliśmy dla wartości progu równej 4% i 5%. Rysunek 4 prezentuje wyniki jakości diagnostycznej dla kolejnych jednostek chorobowych dla trzech modeli. Również w tym przypadku obserwujemy, że model z parametrami wygładzonymi przez parametry Noisy-OR wyznaczone z danych miał najlepsze osiągnięcia.

## 6. Podsumowanie

Jakość diagnostyczna modelu, którego parametry zostały wygładzone na podstawie parametrów Noisy-OR była lepsza o 6.7% od modelu, którego parametry zostały wyznaczone bezpośrednio z danych. Należy zauważyć, że polepszenie jakości diagnostycznej zostało osiągnięte przy niewielkich nakładach, które obejmowały głównie zidentyfikowanie węzłów, które mogą być przybliżone przez bramki Noisy-OR oraz pozyskanie parametrów Noisy-OR z danych czy też od eksperta. Zaobserwowaliśmy również, że jakość diagnostyczna modelu, którego parametry były wygładzone przez parametry pozyskane od eksperta, była porównywalna do modelu, którego parametry były wygładzone przez parametry Noisy-OR wyznaczone z danych. Jakkolwiek jakość diagnostyczna jednostek chorobowych, których mechanizmy nie były dokładnie znane (np. *Hiperbilirubinemia czynnościowa* i *Pierwotna marskość żółciowa*), była lepsza dla modelu wygładzanego przez parametry Noisy-OR wyznaczone z danych niż te pozyskane od eksperta.

Plany na przyszłość obejmują zbadanie właściwości indywidualnych węzłów, tzn. dopasowywanie warunkowych rozkładów prawdopodobieństwa lub Noisy-OR w zależności od węzła.



**Rysunek 4: Jakość diagnostyczna modeli jako funkcja liczby rekordów w jednostce chorobowej**

#### Podziękowania

Praca ta została wykonana w ramach następujących projektów badawczych: grantu KBN 4T11E05522, grantu Air Force Office for Scientific Research F49620-00-1-0112, grantu 501-2-1-02-18/02 Centrum Medycznego Kształcenia Podyplomowego, oraz grantu 16/ST/02 IBIB PAN. Nasza współpraca została wsparta grantem NATO Collaborative Linkage Grant PST.CLG.976167.

Model HEPAR II został utworzony i przetestowany przy użyciu biblioteki klas C++: **SMILE**<sup>®</sup> i narzędzia **GeNIe**, służącego do tworzenia i wnioskowania w graficznych modelach probabilistycznych. Narzędzia powstały w Laboratorium Systemów Decyzyjnych (Decision Systems Laboratory) Uniwersytetu Pittsburgh'skiego i są dostępne na stronie <http://www2.sis.pitt.edu/~genie>.

#### Literatura

- Adlassnig, K. P. i Horak, W. (1995) Development and Retrospective Evaluation of HEPAXPERT-I: A Routinely-used Expert System for Interpretive Analysis of Hepatitis A and B Serologic Findings, *Artificial Intelligence in Medicine*, 7, 1-24.
- Bobrowski L. (1992) HEPAR: Computer system for diagnosis support and data analysis. *Prace IBIB*, 31, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland.
- Cooper G. F. i Herskovits E. (1992) A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9(4):309-347.
- Diez F. J. (1993) Parameter adjustment in Bayes networks. The generalized Noisy-OR gate, *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, 99-105, Washington, D.C., 1993.
- Diez F. J. i Druzdzel M. J. (2002) Canonical probabilistic models for knowledge engineering (w przygotowaniu).

- Diez F. J., Mira J., Iturralde E., i Zubillaga S. (1997), DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, **10**:59-73.
- Henrion M. (1989) Some practical issues in constructing belief networks. W L.N. Kanal, T.S. Levitt, and J.F. Lemmer, edytorzy, *Uncertainty in Artificial Intelligence 3*, 161-173. Elsevier Science Publishers B.V., North Holland.
- Henrion M., Breese J.S. i Horwitz E.J. (1991) Decision analysis and expert systems. *AI Magazine*, **12**(4):64-91.
- Lucas P. J. F. (1994) Refinement of the HEPAR expert system: Tools and techniques. *Artificial Intelligence in Medicine*, **6**:175-188.
- Lucas P. J. F., Segaar R. W., i Janssens A. R. (1989) HEPAR: an expert system for diagnosis of disorders of the liver and biliary tract. *Liver*, **9**:266-275.
- Moore A.W. i Lee M.S. (1994) Efficient algorithms for minimizing cross validation error. W: *Proceedings of the 11th International Conference on Machine Learning*, San Francisco, Morgan Kaufmann.
- Oniśko A., Druzdzel M.J. i Wasyluk H. (2000) Extension of the HEPAR II Model to Multiple-Disorder Diagnosis. *Advances in Soft Computing*, edytorzy: M. Kłopotek, M. Michalewicz, S.T. Wierzhon, 303-313, Physica-Verlag, Heidelberg, New York.
- Oniśko A., Druzdzel M. J., i Wasyluk H. (2001) Learning Bayesian Network Parameters from Small Data Sets: Application of Noisy-OR Gates. *International Journal of Approximate Reasoning*, **27**: 165-182.
- Pearl J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA.
- Pearl J. i Verma T. S. (1991) A theory of inferred causation. W: *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, edytorzy: J.A. Allen, R. Fikes, E. Sandewall, 441-452, Cambridge, MA, Morgan Kaufmann Publishers, Inc., San Mateo, CA.
- Peot M. i Shachter R. (1998) Learning From What You Don't Observe. W: *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, Morgan Kaufmann Publishers, San Francisco, CA, 439-446.
- Richards R. J., Hammitt J. K., i Tsevat J. (1996) Finding the optimal multiple-test strategy using a method analogous to logistic regression. The diagnosis of hepatolenticular degeneration Wilson's disease. *Medical Decision Making*, **16**:367-375.
- Shwe M.A., Middleton B., Heckerman D.E., Henrion M., Horvitz E.J., Lehmann H.P., i Cooper G.F. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, **30**(4):241-255.
- Spirtes P., Glymour C., i Scheines R (1993) *Causation, Prediction, and Search*. Springer Verlag, New York.
- Srinivas S. (1993) A generalization of the Noisy-OR model. W: *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, 208-215, Washington, D.C.
- Wasyluk H. (1995) The four year's experience with HEPAR-computer assisted diagnostic program. W: *Proceedings of the Eighth World Congress on Medical Informatics (MEDINFO-95)*, 1033-1034, Vancouver, BC, July 23-27.