PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1064 — 2005
Pozyskiwanie wiedzy i zarządzanie wiedzą

Marek A. Valenta, Anna Zygmunt
Akademia Górniczo-Hutnicza w Krakowie

POZYSKIWANIE WIEDZY PROBABILISTYCZNEJ DLA MODELU ZAKAŻEŃ SZPITALNYCH

1. Wstęp

Jakość systemów wspomagania decyzji mających charakter systemów ekspertowych zależy od wiedzy zgromadzonej w ich bazach wiedzy. Na potrzeby niniejszej dyskusji zarządzanie wiedzą rozumiane jest jako proces jej pozyskiwania, budowy odpowiedniego modelu wiedzy oraz dobór metody wnioskowania w ścisłym powiązaniu tego procesu z celami tworzenia wynikowego systemu i walidacji tej wiedzy.

W procesie tworzenia baz wiedzy zdecydowanie wyróżnić można dwa jego przypadki: sposób tradycyjny, w którym zasadniczą rolę odgrywa proces pozyskiwania wiedzy od ekspertów i ze źródeł literaturowych, oraz sposób oparty na procesie automatycznym (lub półautomatycznym), w którym wykorzystywane są metody uczenia maszynowego. Ten drugi sposób jest możliwy dzięki istnieniu dziedzinowych operacyjnych baz danych, których dane mogą być podstawą realizacji procesu pozyskiwania wiedzy w sposób automatyczny. Przykładem takiego procesu jest pozyskiwanie wiedzy z baz danych (a nawet hurtowni) o zakażeniach szpitalnych.

W ramach badań prowadzonych przez autorów w Katedrze Informatyki Akademii Górniczo-Hutniczej [1] [5; 18] powstają prototypy modeli wiedzy i systemów ekspertowych, których celem jest wspomaganie podejmowania decyzji związanych

[1] Praca realizowana w Katedrze Informatyki Akademii Górniczo-Hutniczej w Krakowie w ramach projektu badawczego KBN nr 4T11 C023 23, „Konstrukcja modeli wiedzy systemów ekspertowych z uwzględnieniem procesu akwizycji wiedzy i eksploracji baz danych", kierownik projektu: M.A. Valenta.

z problemem występowania stosunkowo dużej liczby przypadków zakażeń szpitalnych. Problem jest bardzo szeroki i samo wspomaganie decyzji może dotyczyć bardzo wielu jego aspektów; nie tylko wspomagania przebiegu procesu terapeutycznego, ale także wspomagania pewnych decyzji organizacyjnych w samych szpitalach oraz wspomagania procesów decyzyjnych związanych z efektywnym zarządzaniem dystrybucją środków finansowych w służbie zdrowia.

Pierwszym, etapem tworzenia systemów takiego typu jest pozyskanie i strukturyzacja wiedzy dla ich baz wiedzy. Zarówno dla lekarzy-ekspertów, jak i w bogatej literaturze przedmiotu naturalnym sposobem reprezentacji wiedzy dotyczącej zagadnień zakażeń szpitalnych jest widzenie rozpatrywanych problemów w kontekście wybranych zdarzeń od siebie zależnych, a zależnościom tym przypisanie pewnych miar łatwo interpretowalnych przez środowisko specjalistów. Nie jest to jedyny możliwy sposób przedstawiania wiedzy z rozpatrywanej dziedziny. Jednak w tym przypadku ze względu na charakter wiedzy dziedzinowej środowisko ekspertów preferuje przedstawianie jej w postaci modeli probabilistycznych.

Jeżeli weźmie się pod uwagę właśnie model probabilistyczny wiedzy jako podstawę tworzenia takich systemów ekspertowych wspomagania decyzji, to przy obecnym stanie możliwości gromadzenia danych o przebiegu i warunkach procesu leczenia w poszczególnych jednostkach służby zdrowia tworzone tam bazy danych mogą się stać cennym źródłem dla procesu automatycznego pozyskiwania odpowiedniej wiedzy. Cennym dlatego, że przy poprawnym doborze źródeł i zakresie tych danych uzyskiwana z nich wiedza winna być w dużym stopniu wiarygodna, co może wpływać zdecydowanie pozytywnie na proces walidacji wiedzy i systemu. Ostatecznie dla realizacji systemów autorzy wybrali dwa rodzaje probabilistycznych modeli wiedzy i odpowiadające im metody wnioskowania: jeden, dopuszczający jedynie bezpośrednie zależności pomiędzy zakażeniami i czynnikami-symptomami z nimi powiązanymi, oraz drugi, bardziej złożony, oparty na sieciach Bayesa.

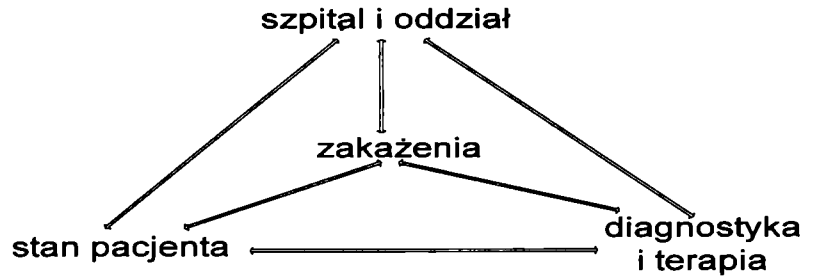
Prowadzone badania generalnie mają na celu ustalenie zakresu stosowalności różnych sposobów akwizycji wiedzy oraz metod i narzędzi realizacji systemów wspomagania decyzji ograniczających ryzyko występowania zakażeń szpitalnych. W badaniach interesujący więc jest nie tylko fakt powstania prototypów systemów czy aspekt techniczny ich realizacji, ale także charakter procesu akwizycji wiedzy i jego dopasowanie do zastosowanego modelu wiedzy oraz zadań i realiów środowiska realizacji tych systemów.

2. Dziedzina, źródła wiedzy i uwarunkowania badań

Dziedzina badań dotyczy szeroko rozumianych zagadnień zakażeń szpitalnych. Są one istotnym problemem, z którym borykają się szpitale na całym świecie [15]. Niezależnie od postępów i rozwoju medycyny, w każdym szpitalu pewien odsetek pacjentów ulega zakażeniu w wyniku samego tam pobytu oraz przeprowadzanego

postępowania medycznego. Przyczyny powstawania zakażeń szpitalnych są bardzo złożone, a problem niezwykle istotny, zwłaszcza z uwagi na skutki tych zakażeń – zarówno zdrowotne [16], jak i kosztowe. W tej sytuacji wspomaganie podejmowania decyzji w zakresie zarządzania procesem terapeutycznym hospitalizowanych pacjentów staje się niemal koniecznością. W dobie rozwoju technologii informatycznych wspomaganie tych procesów przez inteligentne systemy komputerowe wyposażone w bazy wiedzy dziedzinowej wydają się najlepszym rozwiązaniem tego problemu. Może się to wiązać np. z wyznaczaniem przez te systemy preferowanych lub wręcz zalecanych procedur postępowania w konkretnych przypadkach i sytuacjach (standaryzacja procedur medycznych i organizacyjnych), czasem utożsamiane z zarządzaniem wiedzą [17].

Analiza wiedzy z zakresu całości uwarunkowań przebiegu procesów leczenia i ich wpływu na powstawanie zakażeń doprowadziła do logicznej strukturyzacji tej wiedzy [14]. Fakty, z którymi mamy do czynienia w procesie leczenia, można określić mianem czynników wpływających na stopień ryzyka wystąpienia zakażenia. Z punktu widzenia genezy tych czynników ich dokładniejsza analiza wykazuje konieczność zdefiniowania trzech różnych ich grup, są to: czynniki zależne od stanu pacjenta, od ustalonego postępowania diagnostyczno-terapeutycznego oraz czynniki zależne od warunków panujących w szpitalach, a związane z ich infrastrukturą, procedurami organizacyjnymi i poziomem wyszkolenia personelu.



Rys. 1. Struktura wiedzy – grupy czynników ryzyka wystąpienia zakażenia szpitalnego

Celem tworzonych systemów wspomagania decyzji w zakresie przedstawionego problemu jest wspomaganie ustalania, w konkretnych przypadkach, wielkości zagrożenia wystąpienia zakażenia szpitalnego oraz wskazanie czynników ryzyka mających najistotniejszy wpływ na jego wystąpienie. Cele bardziej ambitne, w postaci bezpośredniego wyznaczania optymalnych warunków i przebiegu procesów terapeutycznych minimalizujących ryzyko zakażenia, jakkolwiek pożądane, nie są na razie przedmiotem badań.

Jak we wszystkich dziedzinach, tak i w przypadku zakażeń szpitalnych podstawowym, naturalnym źródłem wiedzy dla przyszłych systemów są eksperci lekarze i literatura przedmiotu. Ogólnie jednak znane są problemy związane z pozyskiwaniem wiedzy heurystycznej z takich właśnie źródeł i jej późniejszą walidacją.

Nie powinien więc dziwić fakt poszukiwania nowych źródeł wiedzy nie tylko przez samych lekarzy, ale i inżynierów wiedzy, tworzących dziedzinowe bazy wiedzy dla przyszłych systemów ekspertowych. Dla inżynierów wiedzy, nie posiadających wiedzy dziedzinowej, jedną z najistotniejszych rzeczy jest, aby alternatywne źródła tej wiedzy były wiarygodne, a użyte metody jej pozyskiwania pozwalały przeprowadzić ten proces poprawnie, mimo braku dogłębnej znajomości problemu. Naturalnie takie dążenia są pożądane, jednak ich całkowita realizacja praktycznie niemożliwa [1; 7].

Na szczęście bardzo wiele szpitali w Polsce wzięło udział w programie monitorowania występowania zakażeń u pacjentów, który to program realizowany był w latach 1999-2001 przez Polskie Towarzystwo Zakażeń Szpitalnych [15]. Jego konsekwencją jest powstanie obszernej bazy danych [12], a ostatnio hurtowni danych [4] zawierającej bezpośrednie wyniki tego programu.

W prowadzonych badaniach właśnie powyższa baza danych i hurtownia stały się źródłem wiedzy dla budowy baz wiedzy przyszłych systemów ekspertowych wspomagania podejmowania decyzji w zakresie oceny ryzyka i zarządzania procesami terapeutycznymi pozwalającymi zmniejszyć ryzyko powstawania zakażeń szpitalnych.

Dane te były pierwotnie i są obecnie wykorzystywane (z użyciem metod statystycznych) przez lekarzy ekspertów do pogłębienia ich wiedzy o zachodzących w szpitalach procesach (np. przy wskazaniu najczęściej występującego typu zakażenia lub wyliczeniu liczby pacjentów ulegających zakażeniu w zależności od różnych czynników ryzyka związanych z tymi zakażeniami).

Dla inżynierów wiedzy rzeczywiste bazy danych przypadków stanowią cenne źródło danych służących do pozyskiwania wiedzy poprzez zastosowanie różnorodnych metod eksploracji danych. Procesy eksploracji baz danych umożliwiają znalezienie w takich danych prawidłowości, zależności i schematów nie tylko znanych ekspertom, ale także tych przez nich „nieuświadomionych". Procesy te, ze względu na rozmiar danych, które należy poddać analizie oraz złożoność tych analiz, realizowane są za pomocą różnych zaawansowanych metod i narzędzi informatycznych [3].

3. Zakres badań i koncepcja rozwiązania problemu

Mając na uwadze różne uwarunkowania akwizycji wiedzy dla budowy probabilistycznych modeli wiedzy oraz definiowania procesów wnioskowania na podstawie modeli, w rozpatrywanym przypadku można stwierdzić przydatność dwóch ich typów.

Pierwszy z modeli wykorzystanych do badań pozwala na realizację podstawowego wnioskowania opartego na prostym klasyfikatorze Bayesa [14]. Baza wiedzy realizowana z wykorzystaniem tego modelu zakłada istnienie dwóch zbiorów zależnych od siebie zdarzeń: zbioru typów występujących zakażeń (hipotez) oraz

zbioru istotnych czynników-symptomów, które stanowią zbiór zdarzeń obserwowalnych. Budowę bazy wiedzy o znacznie większym stopniu złożoności, ale w dalszym ciągu wykorzystującej wiedzę o charakterze probabilistycznym umożliwiają modele oparte na sieciach Bayesa [8; [9; 10].

Pierwszy z systemów implementowano, opierając się na shellu probabilistycznego systemu ekspertowego BayEx [13]. Drugi z modeli, wykorzystujący sieć Bayesa, budowano z wykorzystaniem pakietu BN Power Software [19].

W obu przypadkach z uwagi na dużą złożoność reprezentowanej wiedzy zakres systemu ograniczony został merytorycznie do oceny ryzyka występowania zakażeń ran operacyjnych (ich udział w zakażeniach to tylko około 3%, ale przypadek tych zakażeń jest jednym z lepiej udokumentowanych i jest ciekawy z powodu występowania dużej ilości symptomów mających istotny wpływ na ich występowanie). Tak wybrany problem stanowi więc dobry przykład dla badań, których wyniki winny się dać uogólnić.

Budowa bazy wiedzy dla systemu realizowanego na podstawie shell BayEx to klasyczny przykład pozyskiwania wiedzy bezpośrednio od ekspertów. Eksperci, korzystając ze swojego doświadczenia, wytypowali zbiór istotnych według nich czynników-symptomów i połączyli je z hipotezami zależnościami, którym przyporządkowali wartości prawdopodobieństw. Podstawową weryfikację formalną powstałej bazy wiedzy przeprowadził moduł walidacji wbudowany w BayExa, a w wielu dyskusjach eksperci dodatkowo weryfikowali wymóg metody dotyczący warunkowej niezależności czynników-symptomów. Proces walidacji bazy wiedzy obejmował przeprowadzenie szeregu ekspertyz na „przypadkach testowych" opracowanych przez ekspertów. Pewną istotną niedoskonałością takiej metody jest fakt opracowywania wiedzy systemu i jej walidacji przez ten sam zespół ekspertów. Uzyskiwane w trakcie testowania przebiegi ekspertyz i ich wyniki stały się podstawą do dyskusji i aktualizacji (strojenia) bazy wiedzy. Proces ten okazał się jednak niezwykle trudny z uwagi na brak możliwości obserwowania bezpośredniego wpływu poszczególnych zmian w bazie wiedzy na przebieg (stany przejściowe chwilowej bazy wiedzy) i wyniki ekspertyzy. Ten stan rzeczy stał się istotnym przyczynkiem do sformułowania założeń i budowy systemu xBayEx [11]. Narzędzie to, dodatkowo wizualizując stany chwilowej bazy wiedzy, w znacznym stopniu wspomaga proces walidacji i aktualizacji bazy wiedzy (w założonym modelu), jednak w dalszym ciągu przy pełnym czynnym uczestnictwie ekspertów. Dodatkowo stwierdzono, że problemy z walidacją bazy wiedzy wynikają także ze zbyt uproszczonego modelu wiedzy, w którym eksperci nie mogą uwzględnić znanych im zależności niemających jednak charakteru powiązań hipotezy i czynnika-symptomu.

Powyższe problemy z pozyskiwaniem wiedzy i jej walidacją, a także z uwzględnieniem większej ilości zależności pomiędzy różnymi faktami, nie występują przy kolejnej realizacji systemu ekspertowego, podczas której autorzy wykorzystali model wiedzy w postaci sieci Bayesa, a proces akwizycji oparli na uczeniu

komputerowym. Uzyskano to dzięki realizacji systemu opartego na wspomnianym pakiecie BN Power Software. Wiedza pozyskana w wyniku procesu uczenia realizowanego za pomocą tego oprogramowania może być teoretycznie uznana za bardziej wiarygodną niż w poprzednim przypadku, gdyż powstaje na bazie rzeczywistych danych zaczerpniętych ze wspomnianych baz danych zakażeń szpitalnych [8; 9]. Naturalnie teoretycznie, bo praktyka pokazuje, jak duży wpływ na osiągane wyniki eksploracji danych, czyli też na jakość pozyskanej wiedzy, mają dodatkowe czynności wspomagane przez ekspertów, a wykonywane nawet już na wstępnym etapie eksploracji (wyboru danych czy np. ich dyskretyzacji).

4. Sieci Bayesa, wybrane narzędzia ich realizacji i przykłady uzyskiwanych wyników

Tworzenie modelu wiedzy będącego siecią Bayesa wymaga realizacji dwóch kolejnych etapów: ustalenia struktury sieci, czyli zdefiniowania węzłów (reprezentujących zdarzenia) i powiązań między nimi, oraz kolejno ustalenia wartości mocy tych powiązań. Tak stworzony model, po jego przetestowaniu, może się stać podstawą dla budowy systemu wspomagania decyzji, który w przypadku jego realizacji za pomocą narzędzi BN Power Software określany jest mianem *klasyfikatora*. Proces realizacji takiego zadania oraz schemat modułów funkcjonalnych systemu BN Power Software przedstawione zostały szczegółowo w [19].

Wybór pakietu programowego BN Power Software uwarunkowany był bardzo wysoką jego oceną dokonaną uprzednio przez ekspertów z tej dziedziny podczas KDD Cup 2001 – Data Mining Competition [20]. System BN Power Software ma też ważną cechę, która dopuszcza w działaniu algorytmu automatycznego uczenia się – generowania sieci Bayesa predefiniowanie przez ekspertów wybranych elementów przyszłego modelu wiedzy. Rozbudowany sposób wspomagania przez ekspertów procedur automatycznych pozwala wskazywać na pożądane w modelu związki logiczne oraz ich charakter. Na przykład można wskazywać systemowi dane, pomiędzy którymi istnienie związków logicznych należy z założenia wykluczyć lub jeśli z góry zakładamy, że różnego typu związki między nimi występują (zależności ogólne, związki przyczynowo-skutkowe), a nawet wskazywać na charakter danych uwzględniający ich miejsce w strukturze sieci (korzeń i liść). Rozbudowane narzędzie realizacji procesu budowy modelu pozwala z bardzo dużej ilości danych dostępnych do analizy dokonać wyboru podzbioru danych dla tego procesu istotnych, a dla danych o wartościach ciągłych pozwala wybrać sposób ich dyskretyzacji. Wygodną własnością BN Power Software jest możliwość czerpania danych wejściowych z bardzo zróżnicowanych źródeł danych, a w tym szczególnie z baz danych zgodnych z popularnym formatem MS SQL Server.

Podział systemu BN Power Software na moduły odzwierciedla jego zasadnicze funkcje użytkowe, realizowane z wykorzystaniem różnych metod formalnych. I tak Data Preprocesor pozwala na wstępną selekcję danych wejściowych i ich konwer-

się do postaci ułatwiającej dalsze przetwarzanie. Konwersja ta to przede wszystkim dyskretyzacja wartości realizowana na kilka sposobów predefiniowanych w systemie. BN PowerConstructor to moduł realizujący proces pełnego wygenerowania sieci, oparty na danych wejściowych i predefiniowanych przez użytkownika wybranych jej cechach. Ten model wiedzy dziedzinowej w postaci struktury sieci Bayesa wykorzystuje warunkową niezależność faktów, a uzyskiwany jest metodą przybliżoną z zastosowaniem drzew decyzyjnych [40]. W rozważanym systemie najbardziej zaawansowanym technologicznie modułem jest BN PowerPredictor. Umożliwia on nie tylko wygenerowanie dla wybranej dziedziny wiedzy odpowiadającej jej struktury sieci Bayesa, ale opierając się na wskazanych zbiorach danych wejściowych, potrafi wygenerować tzw. klasyfikator Bayesa. Klasyfikator, wykorzystując kompletną bazę wiedzy w postaci sieci Bayesa (z automatycznie wygenerowaną strukturą i wartościami przypisanymi związkom pomiędzy danymi), może realizować funkcje inteligentnego wspomagania decyzji w dziedzinie zaimplementowanej wiedzy.

W celu przetestowania możliwości w pełni automatycznej generacji wiedzy z baz danych zakażeń szpitalnych [12] zrealizowano kilka kolejnych kroków wyznaczanych przez stosowaną technologię [5]. Pierwszym z nich było odpowiednie przygotowanie danych, zawartych w bazach danych zakażeń szpitalnych, które objęło kolejno:

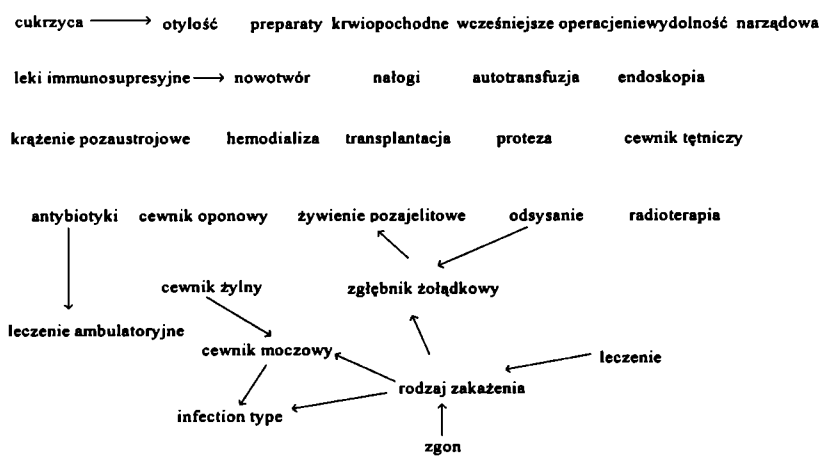
- wybranie tylko tych atrybutów rekordów, które są interesujące do śledzenia w procesie eksploracji i odzwierciedlają najciekawsze aspekty związane z zakażeniami szpitalnymi;
- przekształcenie części danych do postaci wymaganej przez algorytmy BN Power Software (np. dyskretyzacja danych);
- zdefiniowanie akceptowalnych próbek danych, bardzo długie analizy bowiem utrudniają przeprowadzenie symulacji porównawczych dla różnych parametrów algorytmu (zbiór danych zawężono do grupy pacjentów, którzy ulegli zakażeniu rany operacyjnej).

Ważnym elementem przyjętej metodyki badań była selekcja zbiorów danych na dane treningowe, wykorzystywane przez system przy tworzeniu klasyfikatora, i testowe, pozwalające w trakcie procesu walidacji wiedzy na określenie skuteczności uzyskanych klasyfikatorów. Przyjęto założenie, że ¾ danych o interesujących przypadkach, a zgromadzonych w bazie danych stanowi dane treningowe, ¼ zaś to dane testujące, a ich selekcja wykonywana jest na podstawie wbudowanego w system BNPS generatora pseudolosowego.

Dodatkowym zabiegiem metodycznym, wykonanym na potrzeby możliwości merytorycznej walidacji wiedzy i systemu (wygenerowanego klasyfikatora), było wyróżnienie sześciu grup danych – czynników ryzyka, odnoszących się w szczególności do atrybutów związanych z: operacjami, stanem pacjenta, charakterystyką oddziałów oraz szpitali, w których pacjenci rejestrowani w bazie danych byli leczeni.

Automatyczne konstruowanie sieci bayesa na podstawie każdej z grup danych pozwoliło na wygenerowanie kilku jej struktur, które umożliwiły znalezienie interesujących zależności w danych. W ten sposób otrzymano wiedzę dość szczegółową, będącą pewnymi fragmentami wiedzy obejmującej całość zagadnienia zakażeń szpitalnych. Taka wiedza fragmentaryczna jest jednak znacznie łatwiejsza dla jej dyskusji i weryfikacji nie tylko dla inżynierów wiedzy, ale przede wszystkim dla lekarzy ekspertów.

W realizacji badań ważnym elementem algorytmu, mającym duży wpływ na sterowanie istotnością wiedzy, którą chce się pozyskać, jest wartość parametru *próg* (*threshold*). Sterowanie wartością tego parametru umożliwia zmianę ilości i stopnia wykrywanych powiązań. Tak więc niskie wartości parametru (wartości poniżej 1) pozwalają na wyszukiwanie nawet słabych relacji, ale znacznie wydłużają czas działania algorytmu budowy sieci. Z kolei wysokie wartości parametru *próg* (powyżej 1) powodują wyszukiwanie jedynie mocnych powiązań pomiędzy węzłami.

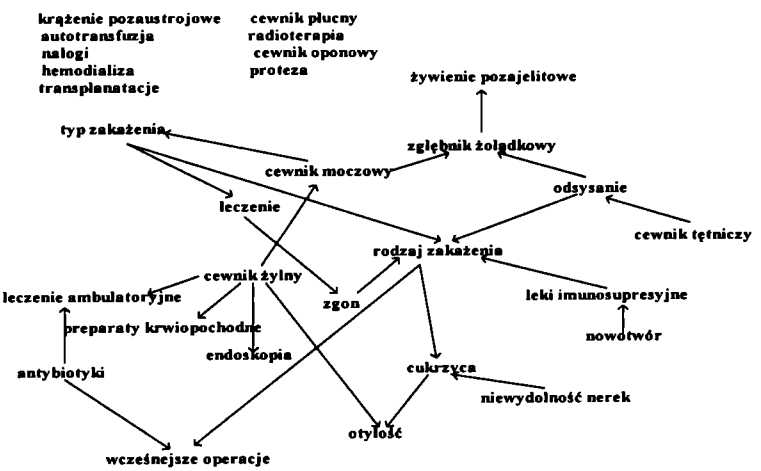


Rys. 2. Sieć wygenerowana dla czynniki_zakażenia – typy_zakażeń dla *progu* = 1.0 [5]

W ramach eksperymentu wygenerowano wspomniane sieci zarówno dla niskich, jak i wysokich wartości *progu*. Przetestowano w ten sposób wrażliwość programu generowania sieci na wartość tego parametru, a także dobrano jego wartości dla różnych wymaganych badaniami modeli. Przykładowo na rys. 2 i 3 pokazano struktury sieci zależności wygenerowane przez moduł BN PowerConstructor ze zbioru takich samych danych wejściowych, a dobranych dla zobrazowania

wpływu różnych *czynników_zakażeń* (czynniki ryzyka występowania zakażenia rany operacyjnej) na *typy_zakażeń* (typy zakażeń związanych z ranami operacyjnymi). Na rys. 3 wyraźnie widoczny jest fakt występowania w wygenerowanej sieci znacznie większej ilości zależności pomiędzy czynnikami, niż ma to miejsce na rys. 2, czyli przy podwyższonej wartości parametru *progu*.

Podobne eksperymenty na danych dotyczących szczegółowych faktów związanych z przebiegiem zabiegu operacyjnego i jego następstw w postaci różnych rodzajów zakażeń doprowadziły do wygenerowania sieci zależności uprawniających do wniosków często pozostających w pewnej sprzeczności z ogólnie przyjętymi zależnościami definiowanymi przez ekspertów lekarzy. I tak np. czas operacji okazał się warunkowo niezależny od typu znieczulenia, wcześniej zastosowanego charakteru leczenia i zastosowanego profilu przeciwbólowego, natomiast jest warunkowo zależny od czystości pola operacyjnego, zastosowanego drenażu czy miejsca operacji. Z kolei typ zakażenia okazał się warunkowo zależny od czystości pola oraz miejsca operowanego, warunkowo niezależny zaś od bardzo wielu pozostałych czynników. Interesująca, lecz kontrowersyjna jest również uzyskana warunkowa zależność zgonu pacjenta. W jednym z otrzymanych modeli zależy on od rodzaju zakażenia, leczenia, czystości pola oraz miejsca operowanego, natomiast nie zależy od techniki operacyjnej czy czasu operacji.



Rys. 3. Sieć wygenerowana dla czynniki_zakażenia – typy_zakażeń dla progu = 0.5 [5]

Jak widać, generalnie metoda dostarcza bardzo dużego materiału do analiz, a możliwość doboru wartości parametru *próg* pozwala dodatkowo badać teoretyczną

istotność otrzymywanej wiedzy. Teoretyczną, bo jednak istnieje absolutna konieczność dyskusji otrzymanej wiedzy z ekspertami. Dyskusja ta ma na celu z jednej strony zwrócenie ich uwagi na wyniki eksperymentu wskazujące na zależności, których dotychczas sobie nie uświadamiali, z drugiej zaś – weryfikację tych fragmentów uzyskanej automatycznie wiedzy, które są niezgodne z ich oczekiwaniami wynikającymi często z wieloletniego doświadczenia.

Powyższe rozważania dobrze charakteryzują etap prac związanych z procesem pozyskiwania wiedzy do budowania struktury logicznej założonego typu modelu. Pozostaje jednak do realizacji jeszcze kolejny etap prac związany z utworzeniem pod kontrolą BN Power Software aplikacji spełniającej funkcję systemu wspomagania podejmowania decyzji, który w tym środowisku określany jest mianem klasyfikatora. Dość zaawansowane są badania metodyczne w zakresie ustalania parametrów prowadzenia procesów generowania klasyfikatorów Bayesa przez pakiet BN PowerPredictor dla różnych klas zadań związanych z ustalaniem ryzyka występowania zakażeń szpitalnych. Pomimo ich trwania, dotychczas uzyskane rezultaty pozwalają autorom na formułowanie wniosków mówiących o dużej przydatności zastosowanego narzędzia i zawartych w nim metod do rozwiązywania postawionego zadania.

5. Zakończenie

Należy sobie zadać pytanie, dlaczego w procesie akwizycji wiedzy z wykorzystaniem metod automatycznego uczenia z gruntu poprawne metody badawcze mogą jednak prowadzić do wygenerowania wiedzy, która będzie niezgodna z rzeczywistą wiedzą dziedzinową. Analiza takich przypadków prowadzi do wniosku, że jednym z najistotniejszych powodów otrzymywania w takim procesie wiedzy nieprawdziwej jest błędne dobranie zbioru danych wejściowych do celu wykonywanej analizy. Istnienie nawet bardzo dużych baz danych przypadków, odgrywających pozytywną rolę w systemach typu transakcyjnego (zarządzania szpitalem, ruchem chorych, kartoteką pacjenta) nie musi być wystarczającym warunkiem do przeprowadzenia na nich procesów eksploracji danych i uzyskania wiedzy praktycznie użytecznej do implementowania inteligentnych systemów decyzyjnych.

Istotnym wynikiem badań jest między innymi wniosek, że budując obecnie systemy gromadzenia danych (typu transakcyjnego), winniśmy – znając powyższe wnioski z eksperymentów – tak projektować te systemy (zakres danych i zapis zależności semantycznych między nimi), aby w przyszłości mogły się stać pełnowartościowym źródłem wiedzy pozyskiwanej metodami automatycznymi.

Przeprowadzone badania wykazują, że przy odpowiednim doborze źródeł danych oraz efektywnym włączeniu ekspertów w kolejne etapy procesu automatycznego pozyskiwania wiedzy można uzyskać wiedzę przewyższającą jakością wiedzę uzyskiwaną jedynie metodami tradycyjnymi. Istniejące i tworzone metody oraz narzędzia realizacji procesu tworzenia modeli wiedzy i aplikacji z nich korzystają-

cych stają się coraz bardziej wygodnymi i wiarygodnymi partnerami w procesie tworzenia nawet złożonych systemów wspomagania decyzji.

Literatura

[1] Anand S.S., Bell D.A., Hughes J.G., *The Role of Domain Knowledge in Data Mining*, Proc. Of the Fourth Int. Conf. on Information and Knowledge Management, USA, 1995.

[2] Chow C.K., Liu C.N., *Approximating Discrete Probability Distributions with Dependence Trees*, IEEE Trans. On Inf. Theo., vol. 14, no. 3, 1968.

[3] Cichosz P., *Systemy uczące się*, Wydawnictwa Naukowo-Techniczne, Warszawa 2000.

[4] Czechowski T., Pyzioł A., *Hurtownia danych – modele danych i ich dopasowanie dla celów realizacji aplikacji*, praca dyplomowa pod kierunkiem M.A. Valenta, w Katedrze Informatyki EAIiE, AGH, Kraków 2004.

[5] Ćwikła J.; *Eksploracja baz danych a modele wiedzy dziedzinowej*, praca dyplomowa pod kierunkiem A. Zygmunt, w Katedrze Informatyki EAIiE, AGH, Kraków 2003.

[6] Han J., Kamber M., *Data Mining – Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[7] Knobbe A., Schipper A., Brockhausen P., *Domain Knowledge and Data Mining Process Decisions*, http://www-ai.cs.uni-dortmund.de/DOKUMENTE/knobbe_etal_2000b.pdf 2000.

[8] Lucas P., Van der Gaag L.C., *Expert Knowledge and its Role in Learning Bayesian Networks in Medicine: an Appraisal*, in AIME 2001, Lecture Notes in Artificial Intelligence 2101, Springer--Verlag, Berlin 2001.

[9] Lucas P., Van der Gaag L.C., Abu-Hanna A., (eds), *Bayesian Models in Medicine*, Proc. European Conf. On Artificial Intelligence in Medicine, AIME'01, Portugal, 2001.

[10] Murphy K., *A Brief Introduction to Graphical Models and Bayesian*, Networks http://www.ai.mit.edu/~murphyk/Bayes/bayes.html 1998.

[11] Stoch D., *Budowa złożonych modeli wiedzy dla realizacji specyfiki zadań*, praca dyplomowa pod kierunkiem M.A. Valenta, w Katedrze Informatyki EAIiE, AGH, Kraków 2004.

[12] Tadeusiewicz R., Majewski J., Zygmunt A. i in., *Raport z realizacji projektu badawczego zamawianego*, KBN nr 2198/P05/98/13, Badania nad opracowaniem ogólnopolskiego modelu zapobiegania i zwalczania zakażeń szpitalnych – Konstrukcja baz danych, kierownik projektu: R. Tadeusiewicz, kierownik tematu: J. Majewski, AGH, Kraków 2001.

[13] Tadeusiewicz R., Valenta M.A. i inn., *Raport z realizacji projektu badawczego zamawianego*, KBN nr 2198/P05/98/13, Badania nad opracowaniem ogólnopolskiego modelu zapobiegania i zwalczania zakażeń szpitalnych – Konstrukcja systemów ekspertowych, kierownik projektu: R. Tadeusiewicz, kierownik tematu: M.A. Valenta, AGH, Kraków 2001.

[14] Valenta M.A., Śnieżyński B., Zygmunt A., *Probabilistyczny model wiedzy o zakażeniach szpitalnych jako podstawa systemu ekspertowego wspomagania rozpoznawania tych zakażeń*, Materiały konferencyjne, V Krajowa Konferencja Modelowanie Cybernetyczne Systemów Biologicznych, Kraków 2000.

[15] Wójkowska J., Żurek I. (red.), *Rejestracja Zakażeń Szpitalnych*, Kontekst, Kraków 1997.

[16] *Zakażenia szpitalne w USA*, http://www.nil.org.pl/xml/nil/gazeta/numery/n1998/n199812/n19981224 – Gazeta Lekarska, 1998.

[17] Zygmunt A., Valenta M.A., *Pozyskiwanie wiedzy z baz i hurtowni danych*, XII Ogólnopolskie Konwersatorium Sztuczna Inteligencja – nowe wyzwania, Siedlce 2001.

[18] Zygmunt A., Valenta M.A., *Wykorzystanie bazy danych zakażeń szpitalnych jako źródła wiedzy o charakterze diagnostycznym*, V Sympozjum MPM 2003 Modelowanie i pomiary w medycynie, Wydawnictwo Katedry Metrologii AGH, Krynica 2003.
[19] http://www.cs.ualberta.ca/~jcheng/bnsoft.htm
[20] http://www.cs.wisc.edu/~dpage/kddcup2001/

PROBABILISTIC KNOWLEDGE ACQUISITION FOR THE NOSOCOMIAL INFECTIONS MODEL

Summary

Experts and their experience are of great importance for building the probabilistic models of domain knowledge, on condition that these experts are able to convert their experience into the knowledge base components. However, in many cases one can find that the domain databases could be very useful. On the basis of that database one can create knowledge model. Such model, as a rule, shouldn't require considerable amount of work for their verification, because it should be automatically correct image of the described reality.

Making such assumption, we will present the process of knowledge acquisition for the nosocomial infections model, where data about the nosocomial infections will be the source of that knowledge. These data were acquired on Polish Society of Hospital Infection initiative from above 100 hospitals in Poland. Within research, in Department of Computer Science AGH the central database, acquiring data from distributed hospital databases, was created.

Nosocomial infection database (storing data from 1999-2000) was used to make research about automatic knowledge acquisition methods for building the probabilistic models based on bayes net. Presented research show the usability of such process on the basis of BN Power Software. Analysis of the following phase of the process show its complexity and conditions having the significant impact on that process success. In the article the exemplary results will be presented. Conclusions indicate the immense impact input data on the created model quality. Theses conclusions also concern the recommendations about the content and structure of the future systems acquiring nosocomial infection data, so one could automatically use them to the probabilistic knowledge model building.