Ewa STRASZECKA
INSTYTUT ELEKTRONIKI, WYDZIAŁ AUTOMATYKI, ELEKTRONIKI I INFORMATYKI, POLITECHNIKA ŚLĄSKA
ul. Akademicka 16, 44-100 Gliwice

Tworzenie reguł diagnostycznych na podstawie danych

Dr hab. inż. Ewa STRASZECKA

Pracuje w Politechnice Śląskiej od 1983 r., obecnie w Zakładzie Elektroniki Biomedycznej Instytutu Elektroniki. Stopień doktora habilitowanego nauk technicznych w zakresie elektroniki otrzymała uchwałą Rady Wydziału Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w 2011 r. Autorka ponad 90 publikacji z zakresu zbiorów rozmytych, teorii Dempstera-Shafera oraz wspomagania diagnozy, m. in. monografii naukowej „Measures of uncertainty and imprecision in medical diagnosis support".

e-mail: ewa.straszecka@polsl.pl

Streszczenie

W pracy przedstawiono metodę tworzenia reguł diagnostycznych o rozmytych przesłankach reprezentujących objawy i nierozmytej konkluzji odpowiadającej diagnozie. Reguły tworzy się na podstawie danych uczących, lecz są one zrozumiałe dla ekspertów i mogą być przez nich weryfikowane. Zbiór reguł dla każdej z diagnoz jest ustalany odrębnie, z zastosowaniem oryginalnego algorytmu eliminacji reguł. Obliczenia dla dwóch benchmarkowych baz danych potwierdzają efektywność proponowanych metod.

Słowa kluczowe: zbiory rozmyte, teoria Dempstera-Shafera, reguły diagnostyczne.

Data-based creation of diagnostic rules

Abstract

A method of diagnostic rule creation is presented in the paper. The rules have fuzzy premises that represent symptoms and a crisp conclusion relevant to the diagnosis. Each rule has an assigned weight that is determined as a value of the basic probability assignment defined in the Dempster-Shafer theory. Having created the rules, there is performed the diagnostic reasoning for a consulted case whose outcomes are values of the *Bel* belief measure (of the Dempster-Shafer theory) for all diagnostic hypotheses. The hypothesis of the maximal belief is the ultimate conclusion. Membership functions of symptoms and the basic probability assignment are found from the training data. Although the rules are created by means of data, they are understandable for human experts who can interpret and verified them. An individual set of rules is provided for each diagnosis. It results from an original elimination algorithm that is proposed in the paper. The elimination process starts from the complete set of rules and the algorithm indicates rule(s) of the lowest diagnostic significance, which are next deleted. Numerical experiments for two benchmark databases show the properties of the proposed method.

Keywords: fuzzy sets, the Dempster-Shafer theory, diagnostic rules.

1. Wstęp

Jednym z istotnych zadań wspomagania diagnozy medycznej jest dostarczenie narzędzi ułatwiających jednoczesną ocenę wielu objawów rozmaitej natury: stwierdzonych na podstawie wywiadu, badania przedmiotowego lub formułowanych w oparciu o pomiary. Ważne jest również, aby spośród wielu możliwych wybrać te procedury diagnostyczne, które w najkrótszym czasie i przy minimalizacji kosztów, nie tylko finansowych, umożliwią postawienie trafnej diagnozy. Narzędzie wspomagania diagnozy powinno więc umożliwiać spójne wnioskowanie dla dowolnej liczby rozmaitych objawów i kilku hipotez diagnostycznych. Wnioskowanie ma być monotoniczne, czyli potwierdzenie kolejnych objawów choroby powinno za sobą pociągać stopniowy wzrost wartości miary potwierdzającej hipotezę o zaistnieniu tej choroby. Powinno być również możliwe wskazanie objawów o największej wartości diagnostycznej na danym etapie diagnozowania. Ponadto wnioskowanie powinno być zrozumiałe dla lekarza, jako przyszłego użytkownika systemu wspomagania diagnozy. Rozwiązanie tych zadań wymaga opracowania reguł o przesłankach reprezentujących objawy i charakteryzowanych przez wagi, odpowiadające ich znaczeniu w procesie potwierdzania lub wykluczania hipotezy diagnostycznej. Reguły takie mogą być formułowane przez ekspertów, a następnie testowane, uzupełniane i dostrajane za pomocą danych uczących [4]. Jednakże, wymagania w stosunku do wspomagania diagnozy rosną, nie tylko w zakresie liczby rozpatrywanych parametrów medycznych [3], ale i czasu, w jakim należy rozwiązywać problem. Należy więc rozwijać metody oparte na wydobywaniu reguł z danych [3, 6]. Reguły te powinny być reprezentowane w sposób czytelny dla ekspertów, w celu ich weryfikacji lub ewentualnego wyznaczenia nowych procedur diagnostycznych [4].

Powyższe zasady budowy reguł diagnostycznych można zachować przy reprezentacji wiedzy diagnostycznej z zastosowaniem teorii Dempstera-Shafera rozszerzonej dla rozmytych elementów ogniskowych. Podstawy proponowanej koncepcji w zastosowaniu do wspomagania diagnozy medycznej zostały przedstawione w publikacjach [8, 9]. Obecnie omawia się wyłącznie zasadnicze elementy koncepcji, które są konieczne do wprowadzenia oznaczeń matematycznych oraz przedstawienia algorytmu konstrukcji zbioru reguł.

Ten ostatni problem jest trudny do rozwiązania, ponieważ generowanie reguł na podstawie danych prowadzi zazwyczaj do wielkiej ich liczby [6], niekorzystnie wpływającej na możliwości generalizacji, praktycznie uniemożliwiając również ich analizę przez ekspertów [3]. W niniejszej pracy zaproponowano algorytm minimalizacji liczby reguł na podstawie przypisywanych im wartości bazowego prawdopodobieństwa, które definiuje się w teorii Dempstera-Shafera. Wartość bazowego prawdopodobieństwa można uważać za wagę reguły, z jaką implikuje ona wybraną hipotezę diagnostyczną. Jak wskazują przykłady obliczeń [8, 9], wagi reguł odnoszących się do tego samego parametru medycznego, lecz do różnych diagnoz, w różny sposób określają wpływ wartości parametru na potwierdzenie hipotezy. Oznacza to, że dla różnych hipotez, reguły odnoszące się do różnorodnych parametrów medycznych mają największą wartość diagnostyczną. Otwiera to drogę to wyznaczenia indywidualnego zbioru reguł dla poszczególnych hipotez w oparciu o dane uczące. W ten sposób można nie tylko zmniejszyć liczbę reguł diagnostycznych, ale i wskazywać parametry, których znajomość może potwierdzić doraźnie dominującą hipotezę lub też wesprzeć hipotezę konkurencyjną.

Niniejsza praca została podzielona na części opisujące kolejno: zasadnicze elementy rozszerzonej teorii Depstera-Shafera i wnioskowanie diagnostyczne w oparciu o nią (pkt 2), eliminację reguł na podstawie wartości bazowego prawdopodobieństwa (pkt 3), przykłady zastosowań dla znanej bazy benchmarkowej Iris i medycznej bazy danych (pkt 4) oraz dyskusję i wnioski (pkt 5). Istotnym uzupełnieniem dla punktu 2 są publikacje [8] i [9], zawierające dokładniejszy opis zasad rozszerzenia teorii Depstera-Shafera na rozmyte elementy ogniskowe.

2. Teoria Dempstera-Shafera i rozmyte elementy ogniskowe

Reguła diagnostyczna ma postać:

$$\text{Jeżeli objaw}_1 \text{ i } \ldots \text{ i objaw}_n \text{ to diagnoza} \quad (\text{z wagą } w) \tag{1}$$

Za objaw uważa się parametr odniesiony do wartości charakteryzujących diagnozę. Odniesienie to ma w ogólnym przypadku postać rozmytą. Rozpatrzmy na przykład diagnozę „przeziębienie" i objaw „gorączka". Ten ostatni rozumie się jako „wysoka temperatura ciała", gdzie parametrem jest X_1 czyli „temperatura ciała"

a wartością (lingwistyczną) „wysoka” interpretowaną przez funkcję przynależności zbioru rozmytego, powiedzmy $\mu(x_1)= \max(0, 1 - 1/\exp(x_1-36))$. Podana interpretacja umożliwia również reprezentację objawu X_2 „kaszel” w postaci funkcji charakterystycznej: $\mu(x_2)= \chi(x_2)=\{1, x_2=$„kaszel”$\}$. W tym przypadku nośnik zbioru rozmytego ogranicza się do pojedynczej wartości „kaszel”. Taki sposób reprezentacji pozwala na jednolitą interpretacją każdego objawu – określonego lingwistycznie lub liczbowo. Za wartość charakteryzująca diagnozę uważa się wartość parametru typową dla diagnozy. Na przykład dla parametru temperatura i diagnozy „przeziębienie” wartością charakteryzującą jest „wysoka”, podczas gdy dla tego samego parametru i diagnozy „zdrowie” wartością tą jest „normalna”. Wartościom charakteryzującym odpowiadają funkcje przynależności zbiorów rozmytych, czyli dla danego parametru określa się tyle funkcji przynależności, ile jest diagnoz. W ten sposób objaw jest formułowany jako „X jest A” gdzie X jest parametrem, a A wartością (lingwistyczną) dla diagnozy D. Jeden lub kilka objawów stanowiących przesłankę reguły (1) jest elementem ogniskowym (prostym lub złożonym), definiowanym w teorii Dempstera-Shafera, dla którego wyznacza się wartość bazowego prawdopodobieństwa. Wartość ta stanowi wagę reguły. Ostatecznie więc proponowana reprezentacja reguły jest następująca:

$$\text{Jeżeli } X_1 \text{ jest } A_r(x_1) \text{ i } \ldots \text{ i } X_n \text{ jest } A_r(x_n) \text{ to } D \quad (m_D(s_r)), \tag{2}$$

Gdzie element ogniskowy $s_r \equiv \{X_1 \text{ jest } A_r(x_1) \text{ i } \ldots \text{ i } X_n \text{ jest } A_r(x_n)\}$, a r oznacza numer reguły dla diagnozy D. Zbiór $S=\{s_r\}$, $r=1,\ldots, |S|$ jest zbiorem elementów ogniskowych dla diagnozy D, a zarazem zbiorem przesłanek reguł, których konkluzją jest D. Ponieważ podczas wnioskowania rozpatruje się kilka diagnoz, potrzebne jest dodanie indeksu oznaczającego numer diagnozy, tak więc rozpatrujemy diagnozę D_l $l=1,\ldots, k$ i elementy ogniskowe s_{rl}, $r_l =1, \ldots, |S_l|$. Funkcje przynależności dla elementów ogniskowych wyznacza się z wykorzystaniem danych uczących [9]. Rozkład bazowego prawdopodobieństwa (RBP) wyznacza się dla zbioru elementów ogniskowych odnoszących się do jednej diagnozy. Musi on spełniać z definicji [2]:

$$m_l(f) = 0; \quad \sum_{\substack{s_{rl} \in S_l \\ \eta_{rl} > \eta_{RBP}}} m_l(s_{rl}) = 1; \tag{3}$$

gdzie f oznacza element fałszywy (brak jakiegokolwiek objawu), a η_{rl} oznacza stopień dopasowania [8] pomiędzy elementem ogniskowym s_{rl}, a przypadkiem danych uczących, wyznaczany w oparciu o obserwowaną u konsultowanego pacjenta wartość parametru i funkcje przynależności objawu. Z (3) wynika, że należy dobrać próg precyzji η_{RBP}, powyżej którego przypadek danych uznaje się za zgodny z elementem ogniskowym (regułą). Znormalizowane częstości zgodności danych uczących z elementami ogniskowymi stanowią wartości RBP [9]. Kiedy reguły są wyznaczone, oblicza się wartość miary przekonania dla badanego przypadku $\boldsymbol{x}^*$ i każdej rozpatrywanych diagnoz $l=1,\ldots, k$ [2, 9]:

$$Bel_l(\boldsymbol{x}^*, \eta_T) = \sum_{\substack{s_{rl} \in S_l \\ \eta^*_{rl} > \eta_T}} m_l(s_{rl}). \tag{4}$$

Próg zgodności η_T, służący do oceny precyzji dopasowania η_{rl}^* pomiędzy regułą r_l, a przypadkiem danych $\mathbf{x}^*$, może, lecz nie musi być równy η_{RBP} [9]. Następnie porównuje się $Bel_l(\boldsymbol{x}^*,\eta_T)$ $l=1,\ldots, k$ i jako ostateczną wybiera się diagnozę o największej wartości Bel. Jeśli nie ma pojedynczej największej wartości, ostateczna diagnoza nie może być wybrana – występuje błąd wnioskowania.

3. Eliminacja reguł

Zachodzi pytanie, które objawy powinny występować w złożonych elementach ogniskowych. Możliwe są następujące rozwiązania: wykorzystanie opinii eksperta, łączenie objawów skorelowanych lub stworzenie wszystkich możliwych kombinacji objawów. Wszystkie te podejścia mają wady: ekspert jest nieosiągalny lub zapis wiedzy jest niedokładny [3], [6], stopień korelacji może znacznie zmieniać się w zależności od danych uczących, a obliczenia dla kompletnego zbioru reguł są czasochłonne i często nie dają najlepszych wyników [10]. W publikacji [10] pokazano, że wartość bazowego prawdopodobieństwa może służyć jako miara wartości diagnostycznej reguły. Można więc stworzyć kompletny zbiór reguł, a następnie eliminować te, dla których wartość bazowego prawdopodobieństwa jest najmniejsza. W niniejszej pracy przeprowadzono eksperymenty polegające na iteracyjnej eliminacji reguł, które mają najmniejszą wartość RBP. W konsekwencji, dla każdej z diagnoz zbiór elementów ogniskowych może być różny, tzn. każda z diagnoz jest stawiana na podstawie indywidualnie wybranego zbioru objawów rozpatrywanych z właściwymi dla niej wagami. Zbliża to proces wspomagania diagnozy do warunków, w jakich decyzję podejmuje ekspert, sprzyja więc tworzeniu reguł, które ekspert potrafi interpretować i weryfikować.

Algorytm wyznaczania reguł diagnostycznych można streścić w następujących punktach.

1. Iteracja=1. Wybrać dane uczące $\boldsymbol{x_i}=[Sx_{i1},\ldots, x_{ic}, D_l]$ $i=1,\ldots, n_l$, $l=1,\ldots, k$, gdzie x_{ij} - wartość j-tego parametru (cechy) dla i-tego przypadku,. Stworzyć funkcje przynależności $\mu_l(x_j)$ $j=1,\ldots, c$.
2. Stworzyć zbiory elementów ogniskowych S_l, $l=1,\ldots, k$.
3. Dla S_l obliczyć RBP m_l, $l=1,\ldots, k$ na podstawie $\boldsymbol{x_i}$, $\mu_l(x_{ij})$ $i=1,\ldots, n_l, j=1,\ldots, c$ i wybranego progu dopasowania η_{RBP}.
4. Wybrać η_T, obliczyć $Bel_l(\boldsymbol{x_i})$ $i=1,\ldots, n_l$, $l=1,\ldots, k$, wybrać diagnozę ostateczną dla $\boldsymbol{x_i}$ i porównać z diagnozą w danych wyznaczając błąd ε, np. jako liczbę błędnych diagnoz.
5. Jeśli iteracja=1 to pkt 7. Jeśli Iteracja>1 to sprawdzić czy ε jest znacząco większy niż w poprzedniej iteracji. Jeśli tak, przywrócić S_l z poprzedniej iteracji i pkt 9, jeśli nie – pkt 6.
6. Sprawdzić, czy w ostatnich dwóch iteracjach otrzymano te same wartości RBP dla k-1 diagnoz. Jeśli tak, to pkt 9, jeśli nie to pkt 7
7. Eliminować reguły poprzez usunięcie z S_l elementów ogniskowych, dla których: $m_l(s_{r_l}) = \min_{l,r_l}\left(m_l(s_{r_l})\right)$ $l=1,\ldots, k$, $r_l=1,\ldots,|S_l|$.

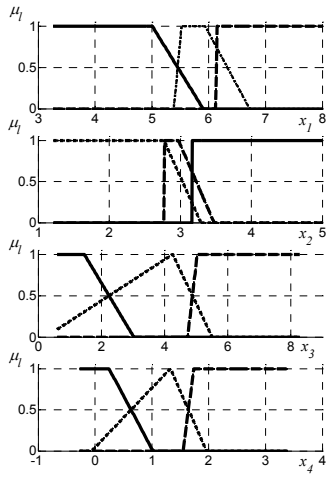
   Jeżeli wynikiem tej eliminacji zbiór pusty, to przywrócić S_l sprzed eliminacji.
8. Przejść do pkt 3.
9. Koniec eliminacji.

Eliminacja reguł postępuje dopóki błąd diagnozy nie zacznie znacząco rosnąć lub RBP nie zaczną się powtarzać dla wszystkich diagnoz oprócz jednej, której element ogniskowy właśnie podlega eliminacji. Jak wskazują eksperymenty (patrz pkt 4), nieznaczne zwiększenie błędu w jednej iteracji nie musi oznaczać, że lepsze rozwiązanie nie może być znalezione. Nawet, jeśli ε=0, należy próbować, czy nie można osiągnąć takiego samego wyniku przy mniejszej liczbie reguł.

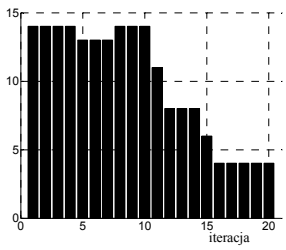
4. Eksperymenty obliczeniowe

Jako przykładowe bazy, dla których stworzono kompletne zbiory reguł i przeprowadzono eliminacje, wybrano bazy z ftp.ics.uci.edu/pub/machine-learning-databases, a mianowicie: bazę Iris oraz bazę new-thyroid, z katalogu thyroid-disease, dotyczącą diagnostyki chorób tarczycy. Obie bazy są benchmarkami. Bazę Iris wybrano, chociaż nie jest ona typowa bazą diagnostyczną, ponieważ jest ona doskonale znana z literatury, co ułatwia czytelnikowi ocenę wyników. Baza new-thyroid dotyczy bezpośrednio zagadnień diagnostycznych, a jej szczegółowe badania opisano w [9]. Obliczenia przeprowadzano używając tego samego zbioru danych (tj. całej bazy) jako danych uczących i testowych, ponieważ celem obecnych badań jest wykazanie, że z pomocą RBP można efektywnie konstruować reguły diagnostyczne, a nie sprawdzenie efektywności metody, którą wykazano w [9]. Wyniki eksperymentów porównano do analogicznych badań z literatury. Uzyskane rezultaty nie wyczerpują możliwości metody, której błąd można zmniejszyć przez dostrajanie RBP dla „trudnych” przypadków [9],

jednak nie zmienia to postaci reguł, dlatego obecnie nie przedstawia się przebiegu ani wyników takiego dostrajania. Jako błąd przyjęto liczbę nieprawidłowo zdiagnozowanych przypadków. Progi η_{RBP} i η_T zmieniano w przedziale [0,1] z krokiem 0.05, a kształty funkcji przynależności modyfikowano, od podstawowego (trapezowego) w kierunku funkcji trójkątnych oraz przeciwnie – do funkcji prawie charakterystycznych [10]. Ostatecznie, przyjęto kształt maksymalnie zmodyfikowany w kierunku trójkąta, jako jeden z możliwych, dających najmniejszy błąd.



Rys. 1. Funkcje przynależności dla bazy Iris: linia ciągła – l=1, linia kropkowana – l=2, linia przerywana – l=3.
Fig. 1. Membership functions for Iris database: solid line – l=1, dotted line – l=2, dashed line – l=3.



Rys. 2. Liczba nieprawidłowych klasyfikacji podczas eliminacji reguł dla bazy Iris
Fig. 2. The number of misclassified cases during the rule elimination for Iris database

Tab. 1. Reguły dla bazy Iris
Tab. 1. Rules for Iris database

Przesłanka (element ogniskowy)	Waga dla klasy		
	r=1 ⇒ kl.1	r=2 ⇒ kl.2	r=3 ⇒ kl.3
X_3 jest $A_r(x_3)$ i X_4 jest $A_r(x_4)$	0,125	0,125	0
X_1 jest $A_r(x_1)$	0,125	0,125	0,145
X_2 jest $A_r(x_2)$	0,125	0,125	0,145
X_3 jest $A_r(x_3)$	0,375	0,375	0,429
X_4 jest $A_r(x_4)$	0,25	0,25	0,281

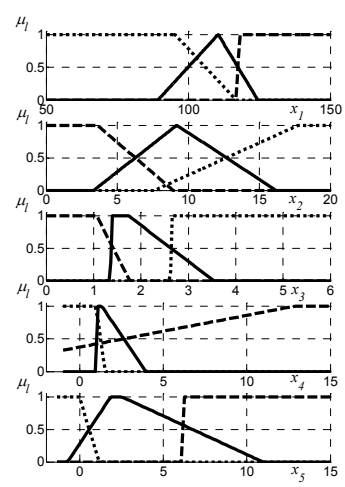
4.1. Baza Iris

Baza Iris zawiera 150 przypadków danych, charakteryzowanych przez 4 cechy (X_1, X_2, X_3, X_4,), należących do 3 klas (Iris-setosa – kl.1, Iris-versicolor – kl.2, Iris-virginica – kl.3), po 50 w każdej klasie. Używając nazewnictwa niniejszej pracy, dla danych określone są 4 parametry i 3 diagnozy. Zostały więc stworzone po 3 funkcje przynależności dla każdego parametru i kompletny zbiór liczący 15 reguł. Wybrano funkcje przynależności pokazane na rys. 1. Najmniejszy błąd otrzymano dla progów η_{RBP}=0.05 i η_T=0.6. W rezultacie eliminacji reguł według algorytmu z pkt.3 uzyskiwano błędy przedstawione na rys.2. Wynik końcowy – 4 błędnie „zdiagnozowane" przypadki, jest nieco lepszy od klasycznej metody grupowania rozmytego (5 błędnych przypadków w [7]) oraz innych metod referencyjnych wykorzystujących zbiory rozmyte (97.3% w stosunku do mniej niż 97% podanych w [1]). Otrzymano go dla reguł zamieszczonych wraz z ich wagami, czyli elementów ogniskowych z ich wartościami RBP, w tab. 1. Zerowa waga oznacza, że element ogniskowy nie należy do zbioru wyznaczonego dla danej klasy.

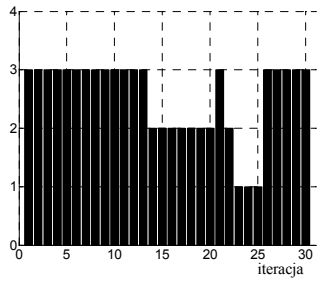
Z tab. 1 wynika, że w efekcie eliminacji otrzymano znacznie mniej reguł, niż zawierały ich zbiory początkowe. Zbiór elementów ogniskowych dla klasy 3 jest inny, niż dla klas 1 i 2, ma również inne wartości RBP, nawet po uwzględnieniu przeskalowania. Liczba reguł (14) jest nieco większa, lecz sumaryczna liczba parametrów występujących w ich przesłankach (16) jest znacznie mniejsza niż w przypadku metody [5], dla której wynoszą one odpowiednio 12 i 38. Proponowana metoda umożliwia więc wyznaczenie indywidualnego zbioru reguł dla każdej z klas, a eliminacja reguł sprzyja poprawie efektywności wnioskowania.

4.2. Baza dla nadczynności tarczycy

Baza danych zawiera przypadki dla 3 diagnoz: nadczynności tarczycy – D_1 (150 przypadków), niedoczynności tarczycy – D_2 (35 przypadków) i eutyreozy, czyli właściwego poziomu hormonów tarczycy – D_3 (30 przypadków). Każdy przypadek charakteryzują wartości 5 parametrów (X_1, ..., X_5). Podobnie, jak dla bazy Iris, w niniejszej pracy dla każdego parametru (cechy) stworzono 3 funkcje przynależności, odpowiednio dopasowując ich kształt. Także dla tej bazy efektywne okazały się funkcje przynależności o kształcie zbliżonym do trójkąta (rys. 3). Stworzono pełny zbiór 31 reguł dla każdej z diagnoz. Z kolei przeprowadzono eliminacje, zgodnie z proponowanym algorytmem. Liczbę nieprawidłowo zdiagnozowanych przypadków podczas eliminacji ilustruje rys. 4. Błędy dla wyznaczonych w proponowany sposób reguł są znacznie mniejsze od uzyskanych dla reguł tworzonych na podstawie korelacji parametrów [8, 9], a nawet przy eliminacji reguł innym sposobem, zaproponowanym w [10]. W trakcie eliminacji błąd diagnozy jednokrotnie nieznacznie wzrósł, ale następnie zmalał poniżej poprzednich wartości. Wynika stąd, że nie należy przerywać eliminacji natychmiast po wzroście błędu. Warto również kontynuować eliminacje nawet, jeśli błąd utrzymuje się na tym samym poziomie, ponieważ można jeszcze zmniejszyć liczbę reguł. Można również zauważyć, że po nadmiernym zmniejszeniu zbioru elementów ogniskowych błąd diagnozy rośnie. Tabela 2 zawiera listę reguł dla wszystkich diagnoz wraz z wartościami RBP uzyskanymi dla η_{RBP}=0.05 i η_T =0.35.



Rys. 3. Funkcje przynależności dla bazy new-thyroid: linia ciągła – l=1, linia kropkowana – l=2, linia przerywana – l=3.
Fig. 3. Membership functions for the new-thyrois database: solid line – l=1, dotted line – l=2, dashed line – l=3



Rys. 4. Liczba nieprawidłowych diagnoz podczas eliminacji reguł dla bazy new-thyroid
Fig. 4. The number of wrong diagnoses during the rule elimination for the new-thyroid database

Tab. 2. Reguły dla bazy new-thyroid
Tab. 2. Rules for the new-thyroid database

Przesłanka (element ogniskowy)	Waga dla klasy		
	r=1 $\Rightarrow D_1$	r=2 $\Rightarrow D_2$	r=3 $\Rightarrow D_3$
X_1 jest $A_r(x_1)$ i X_2 jest $A_r(x_2)$	0,095	0,090	0
X_1 jest $A_r(x_1)$ i X_5 jest $A_r(x_5)$	0,095	0,090	0
X_2 jest $A_r(x_2)$ i X_3 jest $A_r(x_3)$	0,080	0	0,132
X_2 jest $A_r(x_2)$ i X_4 jest $A_r(x_4)$	0	0,087	0,158
X_2 jest $A_r(x_2)$ i X_5 jest $A_r(x_5)$	0,100	0,098	0
X_3 jest $A_r(x_3)$ i X_4 jest $A_r(x_4)$	0	0	0,132
X_3 jest $A_r(x_3)$ i X_5 jest $A_r(x_5)$	0,079	0	0
X_4 jest $A_r(x_4)$ i X_5 jest $A_r(x_5)$	0	0,087	0
X_1 jest $A_r(x_1)$ i X_2 jest $A_r(x_2)$ i X_5 jest $A_r(x_5)$	0,094	0,090	0
X_2 jest $A_r(x_2)$ i X_3 jest $A_r(x_3)$ i X_4 jest $A_r(x_4)$	0	0	0,132
X_2 jest $A_r(x_2)$ i X_3 jest $A_r(x_3)$ i X_5 jest $A_r(x_5)$	0,079	0	0
X_2 jest $A_r(x_2)$ i X_4 jest $A_r(x_4)$ i X_5 jest $A_r(x_5)$	0	0,087	0
X_1 jest $A_r(x_1)$	0,096	0,090	0
X_2 jest $A_r(x_2)$	0,101	0,098	0,158
X_3 jest $A_r(x_3)$	0,080	0	0,132
X_4 jest $A_r(x_4)$	0	0,087	0,158
X_5 jest $A_r(x_5)$	0,101	0,098	0

Liczba reguł dla dwóch diagnoz jest większa, niż w [8] – 9 reguł oraz [10] – 7 reguł, ponieważ wynosi odpowiednio: 11 dla D_1 i D_2, 7 dla D_3. Reguły te jednak są dobrane indywidualnie dla każdej z diagnoz. Należy zwrócić uwagę na fakt, że dla D_3 cechy pierwsza i piąta nie są uwzględniane we wnioskowaniu. Uzyskany błąd jest bardzo mały – tylko 1 przypadek jest źle zdiagnozowany. Jest to wynik lepszy niż w przypadku innych metod wyznaczania zbioru reguł [9], [10]. Ogólna liczba reguł (29), jest taka sama, jak w metodzie [5], ale liczba parametrów występujących w ich przesłankach (52) jest znacznie mniejsza niż dla [5], gdzie wynosi 96.

5. Dyskusja i wnioski

Celem badań było opracowanie metody pozwalającej wyznaczyć zbiór reguł diagnostycznych, które zapewniałyby efektywne wspomaganie diagnozy, a równocześnie mogłyby być interpretowane i modyfikowane przez eksperta. Reguły takie powinny umożliwiać lingwistyczną interpretacje przesłanek, jako objawów, a ich liczba nie może przekraczać możliwości analizy przez eksperta. Można je utworzyć z pomocą teorii Dempstera-Shafera rozszerzonej na rozmyte elementy ogniskowe. W zaproponowanych w pracy regułach przesłanki są zdefiniowane poprzez zbiory rozmyte, które reprezentują objawy a konkluzje stanowią hipotezy diagnostyczne. Przesłanki reguł tworzą zarazem elementy ogniskowe. Wagę reguły we wnioskowaniu określa wartość bazowego rozkładu prawdopodobieństwa przyporządkowana regule.

Maksymalna liczba tak określonych reguł jest równa liczbie wszystkich możliwych kombinacji objawów, a więc zwykle jest zbyt duża. Należy zatem usunąć reguły o małej wartości diagnostycznej, co nie tylko poprawia przejrzystość reprezentacji wiedzy, ale także zwiększa efektywność diagnozy. W artykule na przykładzie dwu eksperymentów pokazano, że wartości bazowego prawdopodobieństwa wyznaczane dla rozmytych elementów ogniskowych mogą stanowić miarę wartości diagnostycznej reguł. Zmniejszając zbiór elementów ogniskowych poprzez eliminację reguł o najmniejszej wartości diagnostycznej można poprawić warunki wspomagania diagnozy równocześnie zmniejszając błąd i liczbę reguł. Jednocześnie wyznacza się zbiory elementów ogniskowych indywidualnie dla każdej z diagnoz, zachowując możliwość interpretacji reguł przez eksperta.

W przedstawionych przykładach konstrukcję reguł rozpoczynano od tworzenia funkcji przynależności dla objawów i sformułowania kompletnych zbiorów elementów ogniskowych dla każdej z diagnoz. Oczywiście, tworzenie tych ostatnich jest praktycznie niemożliwe, jeżeli ma być rozpatrzona wielka liczba parametrów medycznych. Można wtedy rozpocząć obliczenia tworząc reguły dla parametrów nawet słabo skorelowanych, ponieważ nie zauważa się zasadniczej rozbieżności pomiędzy regułami stworzonymi na podstawie korelacji [9] i będącymi wynikiem eliminacji zbioru kompletnego, chociaż w przypadku tych ostatnich błąd jest mniejszy. Jednakże, większe nadzieje można wiązać z postępowaniem odwrotnym do zaproponowanego: uzupełnianiem zbioru reguł o te, które odpowiadają kombinacjom parametrów lub przesłanek reguł o największej wartości diagnostycznej. Niestety, proste „odwrócenie" algorytmu nie daje natychmiast zadowalających rezultatów, ale zagadnieniu temu warto poświęcić uwagę i będzie ono przedmiotem kolejnych prac.

6. Literatura

[1] Duch W., Adamczak R., Grąbczewski K., A new methodology of extraction, optimization and application of crisp and fuzzy logical rules, IEEE Trans. on NeuralNetworks, vol. 12, s. 277-305, 2001.
[2] Dempster A. P., A generalisation of Bayesian inference, J. Royal Stat. Socs. 205-247, 1968.
[3] Kovalerchuk B., Vityaev E., Ruiz J.F., Consistent knowledge discovery in medical diagnosis, IEEE Eng. In Medicine and Biology, vol.19, s.26-37, 2000.
[4] Kwiatkowska M., Atkins A.S., Ayas N.T., Ryan C.F., Integrating knowledge-driven and data-driven approaches for derivation of clinical prediction rules, Proc. ICMLA'05, DO-I:10.1109/ICMLA.2005.41, 6 stron, 2005.
[5] Liu Z., Li Y., A new heyristic algorithm of rules generation based on rough sets, Proc. Int. Seminar on Business and Information Management, IEEE, s.291-294, 2008.
[6] Nozaki K., Ishibuchi H., Tanaka H., Adaptive fuzzy rule-based Classification Systems, IEEE Trans. on Fuzzy Systems, vol.4, no.3, s. 238-250, 1996.
[7] Pedrycz W., Waletzky J.: Fuzzy clustering with partial supervision, IEEE Trans. On Systems, Man and Cybernetics, vol. 27, s.787-795, 1997.
[8] Straszecka E.: Jednoczesna ocena informacji ilościowej i jakościowej podczas wspomagania diagnostyki medycznej. PAK, vol.53, s. 372-375, 2007.
[9] Straszecka E.: Measures of uncertainty and imprecision in medical diagnosis support, Wyd. Politechniki Śląskiej, Gliwice, 2010.
[10] Straszecka E.: The basic probability assignment as a measure of diagnostic rules significance, J. of Medical Informatics & Technologies, Comp. Sys. Dept., Univ. of Silesia, Poland, vol.22, s.95-102, 2013.

otrzymano / received: 07.02.2014
przyjęto do druku / accepted: 03.03.2014 *artykuł recenzowany / revised paper*