Wojciech PODRAZA

**POMORSKA AKADEMIA MEDYCZNA W SZCZECINIE,
KATEDRA I ZAKŁAD FIZYKI MEDYCZNEJ**

Modyfikacja zastosowania teorii zbiorów przybliżonych w medycynie w celu ograniczenia błędów przypadkowych

Dr n. med. Wojciech PODRAZA
– zatrudnienie: Katedra i Zakład Fizyki Medycznej, Pomorska Akademia Medyczna w Szczecinie, stanowisko: adiunkt 1994 r. – obrona pracy doktorskiej w Pomorskiej Akademii Medycznej w Szczecinie, 1983 r. – dyplom lekarza, 1987 r. – I° specjalizacji z pediatrii, 1990 r. – II° specjalizacji z pediatrii.

Streszczenie

Niniejsza praca przedstawia teoretycznie możliwości zastosowania teorii zbiorów przybliżonych w medycynie. Pokazuje, że nieprawidłowa klasyfikacja 1 obiektu (np. błąd przypadkowy) może zaburzyć cały system. Proponuje modyfikację teorii zbiorów przybliżonych w celu uniknięcia błędnej interpretacji wniosków oraz podaje algorytm postępowania.

Abstract

The paper presents theoretical implementation of rough sets theory in medicine. It demonstrates, that improper classification of 1 object (for instance because of an accidental error) may result in destroying of all system. It proposes rough sets theory modification to avoid the results misinterpretation and presents the algorithm of procedure.

Wstęp

Teorię zbiorów przybliżonych opracował Z. Pawlak w 1982 roku [3]. Wielki wkład pracy został włożony w rozwój i zastosowania tej nowej teorii od tego czasu do chwili obecnej. Przegląd rozszerzeń standardowej teorii zbiorów przybliżonych przedstawia praca [11].

Pomimo bardzo dynamicznego rozwoju tej dziedziny wiedzy do chwili obecnej nie ma spektakularnych przykładów jej zastosowań w medycynie. Przeglądając literaturę medyczną sporadycznie można spotkać prace, w których zastosowano teorię zbiorów przybliżonych pisane przez informatyków, matematyków lub lekarzy [2,6,8,9,10]. Nie udało się znaleźć ani jednej tego typu pracy w klinicznej literaturze medycznej.

Podejmowanie decyzji jest jedną z głównych procedur w procesie diagnostyki i leczenia. Powyższe stwierdzenie uzasadnia dalsze wysiłki w celu poszukiwania skutecznych technik wspomagania decyzji w medycynie.

Mogłoby to być zajęcie niewdzięczne z dwóch powodów, z których drugi jest bardziej istotny. Po pierwsze istnieje brak zrozumienia i poparcia tego typu działań w środowisku medycznym. Po drugie nie ma pewności sukcesu. Czy w biologii i medycynie występują charakterystyczne układy wielu cech („szyfry"), które są odpowiedzialne za występowanie określonego wyniku („otwierają zamek szyfrowy") ? Na to pytanie nie można odpowiedzieć na podstawie istniejącego stanu wiedzy. Znalezienie rozwiązanie teoretycznego postawionego pytania jest mało prawdopodobne. Pozostaje praktyczne rozwiązanie tego problemu . Polega ono na napisaniu programu komputerowego w oparciu o teorię zbiorów przybliżonych, wprowadzeniu do programu realnych danych (systemu informacyjnego) oraz analizie tych danych tj. generacji tzw. reguł decyzyjnych. Otrzymanie prostych, jednoznacznych (deterministycznych) reguł decyzyjnych będzie oznaczało sukces, przeciwna sytuacja porażkę. Należy spodziewać się, że tylko w niektórych systemach informacyjnych otrzymamy proste reguły decyzyjne. W praktyce oznacza to, że należy poddać analizie wiele systemów informacyjnych w celu weryfikacji tezy o przydatności teorii zbiorów przybliżonych w medycynie.

Ciekawa, spójna matematycznie, logiczna procedura teorii zbiorów przybliżonych, a także dotychczasowe jej zastosowania pozwalają przypuszczać, że poruszamy się na drodze do sukcesu. Byłby on ukoronowaniem wysiłku intelektualnego naukowców zajmujących się tą problematyką, w tym wielu Polaków.

Teoria zbiorów przybliżonych

System informacyjny

Teoria zbiorów przybliżonych opracowana przez Z. Pawlaka [3,4,5] jest narzędziem analizy danych systemu informacyjnego. Wiele problemów medycznych (klinicznych, biochemicznych i innych) może być przedstawionych w postaci systemu informacyjnego. Przez system informacyjny S rozumiemy:

$$S = < U, Q, V, \rho > \qquad (1)$$

gdzie U jest skończonym zbiorem obiektów, Q skończonym zbiorem atrybutów, $V=\bigcup_{q\in Q} V_q$,
V_q jest dziedziną atrybutu q, $\rho : U \times Q \rightarrow V$ jest funkcją całkowitą, gdzie $\rho (x, q) \in V_q$.

System informacyjny może być utożsamiany z tabelą, w której kolumny odpowiadają atrybutom, a wiersze obiektom systemu. Na przecięciu kolumny q i wiersza x znajduje się wartość ρ (x, q). Funkcja ρ reprezentuje układ wartości atrybutów w tabeli. Wprowadźmy następne pojęcie.

Relacja nierozróżnialności

Niech $S = < U, Q, V, \rho >$ będzie systemem informacyjnym i niech $P \subseteq Q$; $x,y \in U$

Obiekty x i y są nierozróżnialne w systemie S ze względu na podzbiór atrybutów P (oznaczone symbolicznie $x \tilde{P} y$) wtedy i tylko wtedy gdy dla każdego $q \in P$

$$\rho (x,q) = \rho (y,q) \qquad (2)$$

Relacja nierozróżnialności $\tilde{P}$ jest relacją równoważności określoną na zbiorze U. Klasy równoważności relacji $\tilde{P}$ nazywamy P-elementarnymi zbiorami w systemie S.

Przedstawmy powyższe rozważania na przykładzie (tabela 1).

Tab. 1. Przykład systemu informacyjnego

U \ Q	p	q	r	κ
x_1	1	1	1	1
x_2	1	1	1	1
...	1	1	1	1
x_{49}	1	1	1	1
x_{50}	1	1	1	2
x_{51}	2	2	2	1
x_{52}	2	2	2	2
...	2	2	2	2
x_{99}	2	2	2	2
x_{100}	2	2	2	2

System informacyjny stanowi 100 noworodków x_1, ..., x_{100}, które są scharakteryzowane przez atrybuty warunkowe p, q, r (np. płeć, czas trwania ciąży, wyniki badań itp.). Każdy wiersz zawiera wartości poszczególnych atrybutów warunkowych i numer klasy - atrybut konkluzyjny (decyzyjny) κ. Ekspert (lekarz) przypisując numer klasy poszczególnym pacjentom dzieli ich w zależności od potrzeb na np. mających uszkodzenie ośrodkowego układu nerwowego (1) lub zdrowych (2).

Celem całej procedury jest ustalenie związków przyczynowo-skutkowych pomiędzy atrybutami warunkowymi a konkluzyjnymi. Nie zawsze związki te są jednoznaczne i właśnie dlatego zaproponowano teorię zbiorów przybliżonych, która precyzyjnie, matematycznie określa te związki. Niech P = {p, q, r}. P-elementarne zbiory z powyższego przykładu: $X_1 = \{x_1, x_2, ..., x_{50}\}$, $X_2 = \{x_{51}, x_{52}, ..., x_{100}\}$. Wprowadźmy dalsze pojęcia.

Przybliżenie zbiorów w systemie informacyjnym

Niech P* oznacza rodzinę wszystkich klas równoważności relacji określonych na zbiorze U. $Des_p(X)$ oznacza opis klasy równoważności (P-elementarnego zbioru) XP*

$$Des_p(X) = \{(q;v) : \rho(x,q)=v \text{ dla każdego } x \in X \text{ i } q \in P\} \quad (3)$$

W celu scharakteryzowania dowolnego zbioru $Y \subseteq U$ za pomocą zbioru $\{Des_p(X): X \in P^*\}$ Pawlak wprowadził następujące pojęcia:

$$\underline{PY} = \bigcup_{X \in P^* i X \subseteq Y} X \quad (4)$$

P-dolne przybliżenie zbioru Y w systemie S

$$\overline{PY} = \bigcup_{X \in P^* i X \cap Y \neq \emptyset} X \quad (5)$$

P-górne przybliżenie zbioru Y w systemie S

$$Bn_P(Y) = \overline{PY} - \underline{PY} \quad (6)$$

P ograniczenie zbioru Y w systemie S

Dokładność przybliżenia zbioru Y przez zbiór atrybutów P w systemie informacyjnym S zdefiniowano następująco:

$$\mu_P(Y) = \frac{card(PY)}{card(\overline{PY})} \quad (7)$$

gdzie card oznacza liczebność zbioru

$\mu_P(Y)$ jest liczbą z przedziału $<0 ; 1>$. Jeżeli związki przyczynowo-skutkowe pomiędzy atrybutami warunkowymi a konkluzyjnymi są jednoznaczne wówczas $\mu_P(Y) = 1$, w innym przypadku $\mu_P(Y) < 1$.

Klasyfikacja obiektów i jakość klasyfikacji

Niech $\kappa = \{Y_1, ..., Y_n\}$ będzie klasyfikacją obiektów należących do U; $Y_i \cap Y_j = \emptyset$ dla każdego i, j ≤ n, oraz $\bigcup_{i=1}^{n} Y_i = U$. Y_i nazywamy klasami k. W naszym przykładzie $k = \{Y_1, Y_2\}$.

P-dolne i P-górne przybliżenie k w systemie S stanowią odpowiednio zbiory:

$$\underline{P\kappa} = \{\underline{PY_1}, \underline{PY_2}, ... \underline{PY_n}\}$$

$$\overline{P\kappa} = \{\overline{PY_1}, \overline{PY_2}, ... \overline{PY_n}\}$$

Współczynnik $\chi_p(\kappa)$ nazywamy jakością klasyfikacji k przez zbiór atrybutów P i można go obliczyć następująco:

$$\chi_P(\kappa) = \frac{\sum_{i=1}^{n} card(\underline{PY_i})}{card(U)} \quad (8)$$

Tablice i algorytm decyzyjny (konkluzyjny)

Aby przedstawić ideę tworzenia tablic decyzyjnych (konkluzyjnych) podajmy definicję:

zbiór atrybutów $P \subseteq Q$ zależy od zbioru P' w systemie S (zapisujemy symbolicznie P'→P) jeżeli $P' \subset P$

System informacyjny $S=<U, Q, V, \rho>$ możemy przedstawić w postaci:

$$S=<U, C \cup D, V, \rho> \quad (9)$$

$Q=C \cup D$ i $C \cap D=\emptyset$, gdzie C są atrybutami warunkowymi , D atrybutami decyzyjnymi (konkluzyjnymi).

Tak przedstawiony system informacyjny może być traktowany jako tablica decyzyjna (konkluzyjna). Jest ona deterministyczna jeżeli C→D, w przeciwnym razie jest niedeterministyczna.

Niech $D^*=\{Y_1, Y_2, ... Y_n\}$ i $C^*=\{X_1, X_2, ... X_k\}$ wyrażenie $Des_C(X_i) \Rightarrow Des_D(Y_j)$ jest nazywane regułą decyzyjną w S.

Zbiór reguł $\{r_{i,j}\}$ dla każdej klasy Y_j definiuje się następująco:

$$\{r_{i,j}\}=\{Des_C(X_i) \Rightarrow Des_D(Y_j) , X_i \cap Y_j \neq \emptyset, i=(1,2,...k), j=(1,2,...,n)\} \quad (10)$$

Reguła $\{r_{i,j}\}$ jest deterministyczna jeżeli $X_i \cap Y_J = X_i$, w przeciwnym razie jest niedeterministyczna.

Dla podanej wyżej tabeli $\{r_{1,1}\}=\{$ p=1, q=1, r=1 ⇒ κ=1 lub κ=2} reguła niedeterministyczna.

Po eliminacji obiektów x_{50} i x_{51} $\{r_{1,1}\}=\{$ p=1, q=1, r=1 ⇒ κ=1} reguła deterministyczna

Zbiór wszystkich reguł $\{r_{1,1}\}$ nazywamy algorytmem decyzyjnym (konkluzyjnym).

Szczegóły dotyczące algorytmu decyzyjnego przedstawiają prace [1,5,7].

Modyfikacja zastosowania teorii zbiorów przybliżonych

Rozważmy system informacyjny S gdzie Q = {p, q, r, k}(tabela 1).

$X_1 = \{x_1, x_2, ..., x_{50}\}$, $X_2 = \{x_{51}, x_{52}, ..., x_{100}\}$,

$Y_1 = \{x_1, x_2, ..., x_{49}, x_{51}\}$, $Y_2 = \{x_{50}, x_{52}, ...x_{100}\}$

$\underline{PY_1} = 0$ bo $X_1 \not\subset Y_1$ i $X_2 \not\subset Y_1$

$\overline{PY_1} = X_1 \cup X_2$ bo $X_1 \cap Y_1 \neq \emptyset$ i $X_2 \cap Y_1 \neq \emptyset$

$$\mu_P(Y_1) = \frac{card(PY_1)}{card(\overline{PY_1})} = \frac{0}{50+50} = 0$$ (card - oznacza liczebność zbioru)

Analogicznie $\mu_P(Y_2) = 0$

W tym przypadku otrzymujemy informację (reguły konkluzyjne): jeżeli pacjent posiada atrybuty p=1, q=1, r=1 to κ = 1

lub 2, jeżeli posiada atrybuty p=2, q=2, r=2 to $\kappa = 1$ lub 2. Innymi słowy nie uzyskaliśmy żadnej pożytecznej informacji.

Propozycja modyfikacji zastosowania zbiorów przybliżonych dla celów medycznych polega na wyrzuceniu z systemu informacyjnego wszystkich kombinacji n% obiektów (pacjentów) i analizę wszystkich w ten sposób utworzonych nowych systemów informacyjnych. Przyjmuję intuicyjnie, że n powinno znajdować się w przedziale <0 ; 5>. Spośród wielu otrzymanych systemów informacyjnych wnioski wyciągamy z tego, dla którego $\mu_P(Y)$ jest największe, pamiętając o dokonanej modyfikacji.

Można sprawdzić, że w naszym przykładzie systemu informacyjnego S gdzie Q = {p, q, r, κ} (tabela 1), dla którego dotychczas uzyskaliśmy całkowity brak pożytecznej informacji, usunięcie każdej kombinacji 2 % obiektów (2 pacjentów) spowoduje między innymi powstanie systemu informacyjnego bez obiektów x_{50} i x_{51} i w rezultacie

$X_1 = \{x_1, x_2, ..., x_{49}\}$, $X_2 = \{x_{52}, x_{53}, ..., x_{100}\}$.

$Y_1 = \{x_1, x_2, ..., x_{49}\}$, $Y_2 = \{x_{52}, x_{53}, ..., x_{100}\}$

$\underline{PY_1} = X_1$, $\overline{PY_1} = X_1$, $\mu_{1P}(Y_1) = 1$ i analogicznie $\mu_{1P}(Y_2) = 1$

Uzyskujemy jednoznaczne reguły konkluzyjne: jeżeli pacjent posiada atrybuty p=1, q=1, r=1 to $\kappa = 1$ (noworodek ma uszkodzenie OUN), jeżeli posiada atrybuty p=2, q=2, r=2 to $\kappa = 2$ (noworodek zdrowy).

Tab.2. Przykład systemu informacyjnego

U \ Q	p	q	r	κ
x_1	1	1	1	1
x_2	1	1	1	1
x_3	1	1	1	1
x_4	1	1	1	1
x_5	1	1	1	1
x_6	1	1	1	1
x_7	1	1	1	2
x_8	1	1	1	2
x_9	1	1	1	2
x_{10}	1	1	1	2
x_{11}	2	2	2	1
x_{12}	2	2	2	2
x_{13}	2	2	2	2
...	2	2	2	2
x_{100}	2	2	2	2

Jak pokazuje powyższy przykład nieprawidłowa klasyfikacja 2 spośród 100 pacjentów {szumy informacyjne takie jak np. błąd wartości atrybutów warunkowych, błąd klasyfikacji obiektu (pacjenta), wyjątek potwierdzający regułę}burzy cały system.

Algorytm eliminacji obiektów

Tabela 2 przedstawia przykład innego systemu informacyjnego. Oznaczenia są takie jak poprzednio, inne są wartości atrybutów niektórych obiektów.

$X_1 = \{x_1 , x_2, ..., x_{10}\}$, $X_2 = \{x_{11}, x_{12}, ..., x_{100}\}$.

$\underline{Y_1} = \{x_1, x_2, ..., x_6, x_{11}\}$, $\underline{Y_2} = \{x_7, x_8, x_9, x_{10}, x_{12}, x_{13}, ..., x_{100}\}$.

$\overline{PY_1} = X_1 \cup X_2$ ponieważ $Y_1 \cap X_1 \neq \varnothing$ i $Y_1 \cap X_2 \neq \varnothing$, $X_1 \in P^*$ i $X_2 \in P^*$, $\underline{PY_1} = 0$ ponieważ $X_1 \not\subset Y_1$ i $X_2 \not\subset Y_1$

$\mu_P(Y_1) = 0$, $\overline{PY_2} = X_1 \cup X_2$, $\underline{PY_2} = 0$, $\mu_P(Y_2) = 0$

Zastosujmy opisaną powyżej procedurę. Usuńmy wszystkie kombinacje 5% obiektów (w naszym przypadku kombinacje 5 obiektów). Otrzymamy między innymi system informacyjny bez obiektów $x_7, x_8, x_9, x_{10}, x_{11}$ (wyróżnionych w tabeli 2). Wtedy:

$X_1' = \{x_1, x_2, ... x_6\}$, $X_2' = \{x_{12}, x_{13}, ..., x_{100}\}$, $Y_1' = \{x_1, x_2, ..., x_6\}$, $Y_2' = \{x_{12}, x_{13}, ..., x_{100}\}$, $\mu_P(Y_1) = 1$, $\mu_P(Y_2) = 1$

Otrzymaliśmy zdawałoby się satysfakcjonujący wynik. Należy jednak zwrócić uwagę, że pomimo z założenia dającej się zaakceptować 5% redukcji wszystkich obiektów zastosowaliśmy 40% redukcję obiektów należących do zbioru elementarnego X_1 (4 spośród 10 obiektów). W tym przypadku nie można wyciągać rzetelnych wniosków dotyczących obiektów zbioru X_1.

Z powyższych rozważań wynika, że musi być spełniony warunek nieprzekroczenia proporcjonalnej redukcji w każdym zbiorze elementarnym w stosunku do całości. Możemy to przedstawić następująco:

$$U = \sum_{a=1}^{m} X_a \tag{11}$$

$$k_U \leq n\%U \tag{12}$$

gdzie k_U oznacza liczbę wyeliminowanych obiektów, $n \in <0,5>$

$$k_U = \sum_{a=1}^{m} k_{X_a} \tag{13}$$

gdzie k_{x_a} oznacza liczbę wyeliminowanych obiektów ze zbioru X_a

$$k_{x_a} \leq n\% X_a \tag{14}$$

Wnioski

Teoria zbiorów przybliżonych nie znalazła dotychczas szerokiego zastosowania w analizie danych medycznych. Wynika to m. in. z błędnego koła polegającego na tym, że brak spektakularnego efektu zastosowania teorii zbiorów przybliżonych w medycynie ogranicza jej masowe zastosowania, a brak masowego zastosowania dobitnie zmniejsza prawdopodobieństwo uzyskania spektakularnego efektu tj. np. realnego rozwiązania określonego problemu klinicznego.

Teoria zbiorów przybliżonych jest spójna matematycznie i logicznie i wydaje się być odpowiednim narzędziem do rozwiązania niektórych problemów z dziedziny medycyny.

Odpowiednie zdefiniowanie zagadnienia medycznego dla oceny którego chcemy użyć teorii zbiorów przybliżonych ma podstawowe znaczenie.

Wydaje się, że należy prowadzić równoczesne prace nad doskonaleniem teoretycznym opisanej teorii i praktycznym jej zastosowaniem w medycynie.

Literatura

[1] M. BORYCZKA, R. SłOWIŃSKI: Derivation of Optimal Decision Algorithms from Decision Tables Using Rough Sets. Bulletin of the Polish Academy of Sciences, Technical Sciences 1988 , nr 36.

[2] A. OHRN, S. VINTERBO, P. SZYMAŃSKI, J. KOMOROWSKI: Modelling Cardiac Patient Set Residuals Using Rough Sets. Proceedings of American Medical Informatics Association Annual Fall Symp., 1997 (Philadelphia PA).

[3] Z. PAWLAK: Rough Sets. International Journal of Computer Information Sciences 1982, nr 11.

[4] Z. PAWLAK: Rough Classification. International Journal of Man-Machine Studies 1984, nr 20.

[5] Z PAWLAK: Decision Tables and Decision Algorithms. Bulletin of the Polish Academy of Sciences, Technical Sciences 1985, nr 33.

[6] W. PODRAZA, H. PODRAZA: Childhood Leukaemia Relapse Risk Factors. A Rough Sets Approach. Medical Informatics 1999, nr 24 (2).

[7] A. Skowron: Extracting Laws from Decision Tables: A Rough Set Approach. Computational Intelligence 1995, nr 11 (2).

[8] J. STEFANOWSKI, K. SŁOWIŃSKI: Rough Sets as a Tool for Studying Attribute Dependencies in the Urinary Stones Treatment Data Set. W: T.Y. Lin, N. Cercone eds.: Rough Sets and Data Mining. Analysis for Imprecise Data. Kluwer Academic Publishers, Boston / London / Dordrecht 1997.

[9] S. TSUMOTO, H. TANAKA: Incremental Learning of Probabilistic Rules from Clinical Databases Based on Rough Set Theory. Proceedings of American Medical Informatics Association Annual Fall Symposium, 1997 (Philadelphia PA).

[10] A. WAKULICZ DEJA, P. PASZEK: Diagnose Progressive Encephalopathy Applying the Rough Set Theory. International Journal of Medical Informatics, 1997, nr 46 (2).

[11] Y.Y. YAO, S.K.M. WONG, T.Y. LIN: A Review of Rough Set Models. W: T.Y. Lin, N. Cercone eds.: Rough Sets and Data Mining. Analysis for Imprecise Data: Kluwer Academic Publishers, Boston / London / Dordrecht 1997.

Artykuł recenzowany