

XIII Konferencja PLOUG
Kościelisko
Październik 2007

Algorytmy selekcji atrybutów w zadaniach eksploracji danych

Ewelina Szydłowska
Politechnika Opolska, Wydział Elektrotechniki, Automatyki i Informatyki

e.szydłowska@po.opole.pl

1. Wprowadzenie

Eksploracja danych (ang. data mining) jest nauką na pograniczu statystyki, uczenia się maszyn, zarządzania danymi i baz danych, rozpoznawania wzorców, sztucznej inteligencji. Obecnie można znaleźć szereg pozycji literatury dotyczących eksploracji danych, w której często są prezentowane metody i algorytmy, a rzadziej ogólne zasady jak estymacja parametrów czy złożoność obliczeniowa. Eksploracja polega na analizie dużych zbiorów danych obserwacyjnych w celu znalezienia związków i podsumowania danych w oryginalny sposób tak, aby były zarówno zrozumiałe, jak i przydatne dla ich właściciela [HdMh]. Wyniki eksploracji prezentuje się w postaci modeli lub wzorców, które mogą mieć na przykład postać równań liniowych, reguł skupień, drzew czy szeregów czasowych. Eksploracja ma na celu analizę danych obserwacyjnych, czyli takich, które były gromadzone przez pewien okres czasu nie koniecznie na potrzeby eksploracji danych. Eksploracja, określana też mianem odkrywania wiedzy, realizowana jest przeważnie w kilku etapach: wybór docelowych danych, wstępne przetworzenie, przekształcenie danych, poszukiwanie wzorców i zależności, interpretacja odkrytych struktur. Dane wymagają często wcześniejszego przetworzenia na przykład poprzez transformacje (polegają na tworzeniu nowego zbioru cech na podstawie już istniejących) lub wybór podzbioru atrybutów. Selekcja atrybutów, na której skupiono się w artykule polega na wyborze odpowiedniej grupy atrybutów z oryginalnego zbioru, która będzie najlepiej odpowiadała postawionym zadaniom, na przykład klasyfikacji. Metody selekcji atrybutów obejmują między innymi takie techniki jak: redukcja wymiarowości, usuwanie nieistotnych lub nadmiarowych cech, dyskretyzacja pionowa lub pozioma.

2. Problem wielowymiarowości

Zbiór danych, będący zbiorem pobranych pomiarów można przedstawić w postaci kolekcji n obiektów, będących zbiorem p pomiarów. Dane mogą tworzyć różne struktury. Mogą występować w postaci szeregów czasowych. Mogą także tworzyć przestrzenne powiązania, zmuszając tym samym to tworzenia złożonych modeli eksploracji, algorytmów, narzędzi. Dane można zapisać w uproszczonej formie w postaci macierzy o wymiarach $n \times p$. Zbiór danych określany jest też pojęciami dane treningowe, próbka, baza danych. Wiersze tej macierzy w zależności od kontekstu nazywane są: jednostkami, instancjami, encjami, przypadkami, obiektami lub rekordami. Kolumny macierzy określane są mianem zmienne, cechy, atrybuty, pola. Macierzowe zbiory danych mogą łączyć się ze sobą poprzez wystąpienia identycznych wartości w tych samych polach. W ten sposób tworzone są dane wielorelacyjne. Po przez transformacje można je zapisywać w postaci jednej tabeli, jednak jej rozmiar będzie bardzo duży i może to mieć niekorzystny wpływ na późniejsze przetwarzanie danych.

Dane dostępne są najczęściej w jednej z dwóch form: ilościowa lub kategoriowa. Zmienna ilościowa jest wartością numeryczną. Przykładem może być na przykład wiek, dochód. Zmienne kategoriowe przyjmują wartości dyskretne. Przykładem może być płeć, stan cywilny, wykształcenie. Mogą mieć one charakter porządkowy (np.: mało, średnio, dużo) lub symboliczny (np. biały, żółty, czerwony, czarny). W zależności od typu pomiaru stosuje się różne techniki analiz. Dla danych kategoriowych, po zamianie nazw na dyskretne wartości liczbowe, nie stosuje się niektórych obliczeń, na przykład średniej arytmetycznej. Nie można tu także zastosować regresji liniowej, polegającej na przewidywaniu jednej zmiennej jako funkcji innych, którą stosuje się analizach wartości numerycznych.

Przeważnie każdy zbiór danych da się przekształcić do postaci macierzowej. Na przykład dokumenty tekstowe można rozpatrywać jako sekwencję słów i znaków interpunkcyjnych. Jeżeli celem jest opisanie zawartości dokumentu, to można go przedstawić w postaci macierzowej. W takiej formie wiersze mogą reprezentować dokumenty, a kolumny słowa. Wartości pól mogą okre-

ślać ilość powtórzeń wyrazu w danym dokumencie. Można przyjąć także zapis 0-1, gdzie 0 oznacza brak wyrazu w dokumencie, a 1 wystąpienie wyrazu. Innym przykładem są dane transakcyjne takie jak rejestr zakupów, dziennik logowań itp. Tutaj również można zastosować zapis macierzowy, przedstawiając w postaci wierszy poszczególne transakcje (osoby kupujące, logowania na stornach WWW), a w postaci kolumn szczegóły transakcji (zakupione produkty, adresy stron WWW). Niestety taki zapis nie przechowuje informacji sekwencyjnych czy czasowych (kolejność zakupów, kolejność odwiedzin stron WWW).

Analiza zbiorów danych w wyniku procesu eksploracji prowadzi do utworzenia modeli i wzorców. Model ma charakter globalny, niesie informację o każdym obiekcie w przestrzeni pomiarowej. Na przykład model liniowy można przedstawić w postaci równania $Y=ax+b$, gdzie Y , X są zmiennymi, a a , b parametrami modelu uzyskanymi w procesie eksploracji. Wzorzec ma charakter lokalny, dotyczy określonego obszaru przestrzeni pomiarowej. Wyrażane jest ograniczeniami nałożonymi na zmienne X, Y np. $X > x_1$, $Y > y_1$. Metody wykrywania wzorców lokalnych stosuje się często w wykrywaniu nieprawidłowości na przykład usterek, fałszerstw bankowych itp.

Eksploracja danych opiera się często na poszukiwaniu modeli i wzorców w zagadnieniach wielowymiarowych. W zadaniach o wysokim poziomie dokładności ilość potrzebnych danych wzrasta wykładniczo wraz z wymiarowością. Takie zachowanie określane jest mianem „przekleństwa wymiarowości”. Wielowymiarowość zależy od złożoności modelu i ilości dostępnych danych. Może się odnosić zarówno do tak małej liczby jak 10 lub tak dużej jak 1000. Często nie wszystkie spośród p zmiennych X są konieczne do dokładnego przewidywania zmiennych wynikowych Y , a niektóre mogą być zupełnie niezwiązane z wyjściem. Na przykład w eksploracji zagadnień finansowych miesiąc urodzenia osoby nie powinien mieć wpływu na jej wypłacalność. Istnieją także zmienne nadmiarowe, to znaczy takie, które powtarzają informacje zawarte w innych zmiennych. Są to przeważnie zmienne wysoko ze sobą skorelowane (na przykład cena netto i brutto). Problem wielu zmiennych można rozwiązać na dwa sposoby:

- Wybór zmiennych, czyli znalezienie podzbioru p' zmiennych, gdzie $p' \ll p$. Istotność zmiennych można oceniać na przykład w sposób ilościowy analizując niezależność:
 - Jeśli $p(y|x_1)=p(y)$ to zmienna wyjściowa Y jest niezależna od zmiennej wejściowej X_1
 - Jeśli $p(y|x_1, x_2)=p(y|x_2)$ to zmienna wyjściowa Y jest niezależna od zmiennej wejściowej X_1 jeżeli znana jest wartość zmiennej wejściowej X_2 .

W praktyce wybór zmiennych najczęściej odbywa się na podstawie szacowania i ocenie stopnia zależności. Zmienne można na przykład szeregować biorąc pod uwagę współczynnik korelacji X z Y . W przypadku zmiennych kategoriowych można określić średnią wzajemną informację między Y a X' (X' może być zmienną kategoriową lub wartością dyskretyzowaną zmiennej X):

$$I(Y; X') = \sum_{i,j} p(y_i, x'_j) \log \frac{p(y_i, x'_j)}{p(y_i)p(x'_j)}$$

Zależność pojedynczych zmiennych X i Y nie informuje o tym jak na jakość modelu eksploracji może wpływać większy zbiór zmiennych. Wybranie k najlepszych pojedynczych zmiennych X i najlepszy zbiór o rozmiarze k zmiennych są zupełnie różnymi zadaniami i należą do zbioru $2^p - 1$ rozwiązań. Sprawdzenie wszystkich możliwości jest bardzo trudne, zwłaszcza przy dużej liczbie p . W praktyce wybieranie podzbiorów odbywa się heurystycznie selekcji, poprzez dodawanie lub usuwanie jednej zmiennej w danym momencie.

- Przekształcenie p oryginalnych zmiennych X na nowy zbiór p' zmiennych Z , gdzie $p' \ll p$; Przykładem może być analiza składowych głównych. Polega ona na tworzeniu nowych

zmiennych na podstawie liniowych kombinacji oryginalnych zmiennych. W tym przypadku zmienne Z są zdefiniowane jako funkcje oryginalnych zmiennych X . Zmienne te są często nazywane funkcjami bazowymi, czynnikami, zmiennymi utajonymi, składowymi głównymi.

3. Selekcja atrybutów z zastosowaniem teorii zbiorów przybliżonych (ang. Rough sets)

Teoria zbiorów przybliżonych dostarcza metody do określenia najważniejszych atrybutów systemu informacyjnego bez utraty zdolności klasyfikacji (lub z małą różnicą) w porównaniu z oryginalnym zbiorem atrybutów. Została wprowadzona przez Zdzisława Pawłaka w 1982 roku. Oparta jest ona na koncepcji górnej i dolnej aproksymacji zbioru, przestrzeni aproksymującej i modeli zbiorów.

Dane przedstawia się w postaci tablicy, w której kolumny odpowiadają atrybutom, a wiersze odpowiadają obiektom. Zapis taki nosi nazwę systemu informacyjnego, który formalnie zapisuje się jako [WaMa99, Dom04]:

$$SI = (U, A, V, f) \quad (1)$$

gdzie:

- U jest niepustym, skończonym zbiorem nazywanym uniwersum; elementy zbioru $\{x_1, x_2, \dots, x_n\}$ U nazywamy obiektami;
- A jest niepustym, skończonym zbiorem atrybutów; system może mieć postać tablicy decyzyjnej w której $A = C \cup D$, gdzie C jest zbiorem atrybutów warunków, a D zbiorem atrybutów decyzyjnych;
- $V = \bigcup_{a \in A} V_a$ jest dziedziną atrybutów, gdzie V_a jest dziedziną atrybutu $a \in A$
- f jest funkcją decyzyjną (funkcją informacji), gdzie $f: U \times A \rightarrow V$

Pary obiektów, które posiadają takie same wartości dla wszystkich atrybutów ze zbioru $B \in A$ określają relację nierozróżnialności (ang. indiscernibility relation) na zbiorze U definiowaną jako:

$$IND(B) = \{x_i, x_j \in U : \forall b \in B, f(x_i, b) = f(x_j, b)\} \quad (2)$$

Każda relacja nierozróżnialności dzieli zbiór na rodzinę rozłącznych podzbiorów zwanych klasami abstrakcji (równoważności) lub zbiorami elementarnymi. Poszczególne klasy nazywamy zbiorami B - elementarnymi i oznaczamy przez $U/IND(B)$. Klasy tej relacji zawierające obiekt x_i oznaczamy $[x_i]_{IND(B)}$. Zbiór $[x_i]_{IND(B)}$ zawiera więc te wszystkie obiekty systemu SI , które są nierozróżnialne z obiektem x_i względem zbioru atrybutów B . Wszystkie elementy każdego zbioru B - elementarnego mają te same wartości wszystkich atrybutów należących do zbioru B (nie są względem nich rozróżnialne). Konstrukcja zbiorów elementarnych jest pierwszym krokiem w klasyfikacjach przy użyciu zbiorów przybliżonych.

Metoda zbiorów przybliżonych opiera się na zagadnieniu aproksymacji. Zbiór $X \subseteq U$ jest zbiorem B - przybliżonym, gdy nie jest skończoną sumą zbiorów B -elementarnych. Aproksymacja dolna oznacza, że elementy bez wątpliwości należą do zbioru:

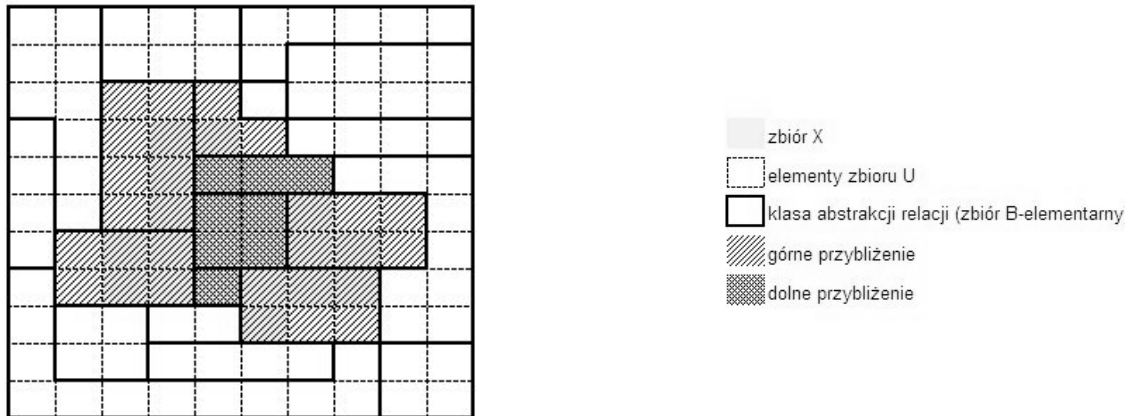
$$\underline{B}X = \{x_i \in U \mid [x_i]_{IND(B)} \subset X\} \quad (3)$$

Aproksymacja górna mówi, że elementy mogą należeć do zbioru. Można ją zdefiniować następująco:

$$\overline{BX} = \{x_i \in U \mid [x_i]_{Ind(B)} \cap X \neq \emptyset\} \quad (3)$$

Granicą (brzegiem) (ang. boundary) w systemie SI nazywamy zbiór będący różnicą górnej i dolnej aproksymacji:

$$BNX = \overline{BX} - \underline{BX} \quad (3)$$



Rys. 1. Graficzna interpretacja aproksymacji

Podstawą teorii zbiorów przybliżonych jest redukt (ang. reduct). Jest to zasadnicza część systemu informacyjnego, która pozwala na rozróżnienie wszystkich rozróżnialnych obiektów z oryginalnego zbioru atrybutów. Ważnym pojęciem jest także rdzeń, będący częścią wspólną wszystkich reduktów.

Postawione zadania polegają na redukcji liczby atrybutów i obserwacji jak ona wpływa na rozróżnialność zbiorów. Jeżeli przy usunięciu atrybutu liczba zbiorów elementarnych się nie zmniejsza to oznacza, że atrybut był zbyteczny. Rozważmy zbiór 10 obiektów, z których każdy jest opisany przez 3 atrybuty (tabela 1). Jeśli weźmiemy pod uwagę wszystkie trzy atrybuty to utworzymy pięć zbiorów elementarnych (tabela 2). Jeżeli weźmiemy pod uwagę przestrzeń atrybutów $B_1=\{a_1, a_2\}$ lub $B_2=\{a_1, a_3\}$ to otrzymamy 4 zbiory elementarne. Jeżeli jednak weźmiemy pod uwagę przestrzeń $B_3=\{a_2, a_3\}$ to otrzymamy 5 zbiorów elementarnych, czyli tyle samo, co w oryginalnej przestrzeni atrybutów A. Oznacza to, że atrybut a_1 jest nadmiarowy, natomiast atrybuty a_2 i a_3 są nieodzowne.

Tabela 1. Zbiór danych

U
 a_1
 a_2
 a_3
 x_1
 2
 1
 3

U/A
 a_1
 a_2
 a_3
 x_1, x_3, x_9

Tabela 2. Zbiory elementarne

Nr zbioru

1

		2	
x ₂		1	
3		3	
2			
1			2
		x ₂ ,x ₇ ,x ₁₀	
x ₃		3	
2		2	
1		1	
3			3
		x ₄	
x ₄		2	
2		2	
2		2	
3		3	
			4
x ₅		x ₅ ,x ₈	
1		1	
1		1	
4		4	
			5
x ₆		x ₆	
1		1	
1		1	
2		2	
x ₇		1	
3		2	
2			
1			
x ₈			
1			
1			
4			
x ₉			
2			
1			
3			
x ₁₀			
3			
2			
1			

Pierwszym etapem w wyznaczeniu reduktów i rdzenia jest utworzenie macierzy rozróżnialności (ang. discernibility matrix). Jest to macierz o rozmiarach $n \times n$, gdzie n jest liczbą zbiorów elementarnych. Elementami macierzy są zbiory atrybutów rozróżniające zbiory elementarne i oraz j . Ponieważ z założenia każdy zbiór elementarny jest inny to macierz nie będzie posiadała pustych komórek. Macierz rozróżnialności jest symetryczna, więc wystarczy rozpatrywać jej dolną lub górną diagonalną część. Na przykład, jeśli dwa zbiory elementarne (np. 1 i 4) z tabeli 2 będą różniły się dwoma atrybutami, to elementem macierzy rozróżnialności będzie zbiór dwuelementowy $\{a_1, a_3\}$. Macierz rozróżnialności dla przypadku z tabeli 2 przedstawiono w tabeli 3.

Tabela 3. Macierz rozróżnialności dla SI z tabeli 1

1
2
3
4
5
1
-
2
a_1, a_2, a_3
-
3
a_2
a_1, a_3
-
4
a_1, a_3
a_1, a_2, a_3
a_1, a_2, a_3
-

Tabela 4. Uproszczony system informacyjny

U/A	Nr zbioru
	1
x_1, x_3, x_9	
1	
3	
	2
x_2, x_7, x_{10}	
2	
1	
	3
x_4	
2	
3	
	4
x_5, x_8	
1	
4	
	5
x_6	
1	
2	

$$\begin{array}{c}
 5 \\
 a_1, a_3 \\
 a_1, a_2, a_3 \\
 a_1, a_2, a_3 \\
 a_3 \\
 -
 \end{array}$$

W następnym kroku należy wyznaczyć funkcję rozróżnialności. Jest to funkcja boolowska skonstruowana w następujący sposób: każdemu atrybutowi ze zbioru atrybutów, który rozróżnia dwa zbiory elementarne (np. $\{a_1, a_2, a_3\}$) przypisujemy zmienną boolowską 'a'. Funkcja przyjmuje wtedy postać $(a_1 + a_2 + a_3)$ (lub $(a_1 \vee a_2 \vee a_3)$). Jeżeli zbiór atrybutów jest pusty to przypisujemy funkcji wartość 1. Macierz rozróżnialności z tabeli 5 funkcja będzie miała postać:

$$f(A) = (a_1 + a_2 + a_3)a_2(a_1 + a_3)(a_1 + a_3)(a_1 + a_3) (a_1 + a_2 + a_3) (a_1 + a_2 + a_3)(a_1 + a_2 + a_3) (a_1 + a_2 + a_3)a_3 = a_2a_3$$

Znalezienie rozwiązania polega na minimalizacji funkcji rozróżnialności. W najprostszej postaci można skorzystać z prawa absorpcji według którego $a \vee (a \wedge b) = a$ i $a \wedge (a \vee b) = a$. Jeżeli, w analizowanym przykładzie, zbiór elementarny nr 1 różni się od zbioru elementarnego nr 2 wszystkimi atrybutami, a od zbioru nr 3 tylko jednym atrybutem, to wystarczy wziąć do rozważań atrybut a_2 : $(a_1 + a_2 + a_3)a_2 = a_2$.

Zapis uproszczonej funkcji można przedstawić w postaci rozłącznej (np. $a_1a_2 + a_2a_3$) lub łącznej (np. $a_2(a_1 + a_3)$). Postać rozłączna, będąca alternatywną reprezentacją systemu informacyjnego SI, wskazuje możliwe redukt. W analizowanym przykładzie funkcja rozróżnialności będzie miała postać $f(A) = a_2a_3$. Oznacza to, że w rozważanym zbiorze atrybutów i obiektów został odnaleziony tylko jeden redukt. Zatem atrybuty a_2 i a_3 dostarczają taką samą wiedzę o uniwersum U jak zbiór atrybutów a_1, a_2, a_3 (tabela 4).

Teorię zbiorów przybliżonych można wykorzystać także w zadaniach klasyfikacji. Wtedy system informacyjny ma postać tablicy decyzyjnej, a atrybuty dzielą się na dwie grupy: warunkowe i decyzyjne (tabela 5). Atrybut decyzyjny określa przynależność obiektów do klas. W analizowanym przykładzie będą to trzy klasy $\{x_1, x_3, x_9\}$, $\{x_2, x_4, x_7, x_{10}\}$ i $\{x_5, x_6, x_8\}$. Zbiór reduktów wyznaczy jest tutaj w bardzo podobny sposób. Macierz rozróżnialności nie jest budowana w oparciu o zbiory elementarne, a biorąc pod uwagę wszystkie obiekty. Pola macierzy rozróżnialności są zbiorami atrybutów, które rozróżniają dwa obiekty x_i i x_j , przy czym obiekty nie należą do tej samej klasy (określonej przez atrybut decyzyjny). Na przykład, w analizowanym zadaniu, obiekty x_1 , x_3 i x_9 należą do tej samej klasy. W związku z tym nie będą one porównywane w macierzy rozróżnialności. Dla tak zbudowanej macierzy funkcję rozróżnialności będzie miała postać:

$$\begin{aligned}
 f(D) &= (a_1 + a_2 + a_3)a_2(a_1 + a_3)(a_1 + a_3)(a_1 + a_2 + a_3)(a_1 + a_3) (a_1 + a_2 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3) \\
 &x(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)a_2(a_1 + a_3)(a_1 + a_3)(a_1 + a_2 + a_3)(a_1 + a_3) (a_1 + a_2 + a_3) \\
 &x(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)a_2(a_1 + a_3)(a_1 + a_2 + a_3)(a_1 + a_3) (a_1 + a_2 + a_3) \\
 &x(a_1 + a_2 + a_3)(a_1 + a_3) (a_1 + a_2 + a_3)x(a_1 + a_2 + a_3)(a_1 + a_2 + a_3) (a_1 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3) \\
 &= a_2(a_1 + a_3) = a_1a_2 + a_2a_3
 \end{aligned}$$

Funkcja posiada dwa redukt $\{a_1, a_2\}$ i $\{a_2, a_3\}$. Oznacza to, że tabelę decyzyjną można uprościć formy przedstawionej w tabelach 6 i 7.

Tabela 5. Tablica decyzyjna

U
a₁
a₂
a₃
d
x ₁
2
1
3
1
x ₂
3
2
1
2
x ₃
2
1
3
2
x ₄
2
2
3
2
x ₅
1
1
4
3
x ₆
1
1
2
3

Tabela 6. Tablica decyzyjna dla reduktu {a₁,a₂}

U
a₁
a₂
d
x ₁
2
1
1
x ₂
3
2
2
x ₃
2
1
1
x ₄
2
2
2
x ₅
1
1
3
x ₆
1
1
3
x ₇
3
2
2
x ₈
1

Tabela 7. Tablica decyzyjna dla reduktu {a₂,a₃}

U
a₂
a₃
d
x ₁
1
3
1
x ₂
2
1
2
x ₃
1
3
1
x ₄
2
3
2
x ₅
1
4
3
x ₆
1
2
3
x ₇
2
1
2
x ₈
1

x ₇	1	4
3	3	3
2		
1	x ₉	x ₉
2	2	1
	1	3
x ₈	1	1
1		
1	x ₁₀	x ₁₀
4	3	2
3	2	1
	2	2
x ₉		
2		
1		
3		
1		
x ₁₀		
3		
2		
1		
2		

Tabela 8. Macierz rozróżnialności dla tablicy decyzyjnej z tabeli 5

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 1
-

2

 a_1, a_2, a_3

-

3

-

 a_1, a_2, a_3

-

4

 a_2

-

 a_2

5

 a_1, a_3 a_1, a_2, a_3 a_1, a_3 a_1, a_2, a_3

6

a_1, a_3
 a_1, a_2, a_3
 a_1, a_3
 a_1, a_2, a_3
 -
 -

7

a_1, a_2, a_3
 -
 a_1, a_2, a_3
 -
 a_1, a_2, a_3
 a_1, a_2, a_3
 -

8

a_1, a_3
 a_1, a_2, a_3
 a_1, a_3
 a_1, a_2, a_3
 -
 -
 a_1, a_2, a_3
 -

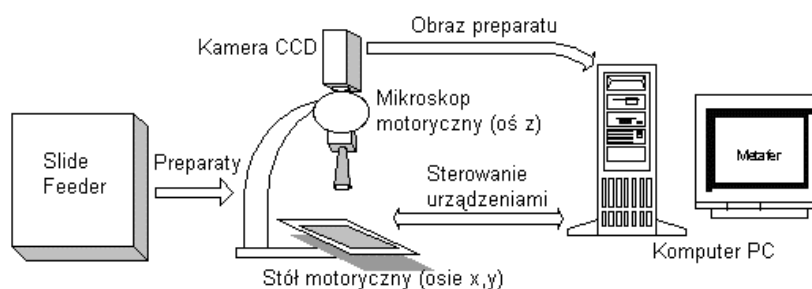
9

-
 a_1, a_2, a_3
 -
 a_2
 a_1, a_3
 a_1, a_3
 a_1, a_2, a_3
 a_1, a_3
 -

10
 a_1, a_2, a_3
 -
 a_1, a_2, a_3
 -
 a_1, a_2, a_3
 a_1, a_2, a_3
 -
 a_1, a_2, a_3
 a_1, a_2, a_3
 -

4. Zbiory przybliżone w zagadnieniach wykrywania komórek rakowych

Walka z nowotworami jest jednym z głównych problemów medycyny. Szanse na przezwyciężenie choroby są znacznie większe jeżeli zostanie ona wykryta we wczesnym stadium. Taką możliwość dają techniki mikroskopowe jaką jest na przykład analiza próbek fluorescencyjnej hybrydyzacji in situ (ang. Fluorescence in situ hybridization, FISH) [BCBK03, Bro01]. Odbywa się ona w środowisku laboratoryjnym z zastosowaniem systemów komputerowych wyposażonych w oprogramowanie do analizy cytogenetycznej. Znaczącymi elementami systemu są: mikroskop, kamera CCD oraz stół skaningowy (rys. 2). Badania polegają na pobraniu obrazu CCD z odpowiednio przygotowanej próbki i analizie cech morfometrycznych wyodrębnionych na próbce obiektów. Wyniki analizy dostępne są w postaci plików tekstowych zawierających zbiór wartości wybranych parametrów. Szczegółową budowę stanowiska laboratoryjnego oraz sposoby wykonywania pomiarów można znaleźć w literaturze [PILo01, Guz05].



Rys. 1. Komponenty systemu "Metafer" [Guz05]

Klasyfikacja komórek pęcherza moczowego została przeprowadzona w oparciu o algorytmy eksploracji danych. Dane pochodzące z systemu skaningowego „Metafer” zawierały ocenione przez eksperta obrazy pól przebadanych preparatów. Wyodrębnione obiekty zostały podzielone na dwie grupy: komórki rakowe i komórki zdrowe. Każdy obiekt scharakteryzowany jest poprzez 217 atrybutów opisujących morfometryczne cechy komórki, jak rozmiar, kształt itd. Analizowany zbiór zawierał 22962 obiektów wśród których 640 zostało zidentyfikowanych przez eksperta jako komórki rakowe. W zadaniu eksploracji danych niektóre cechy mogą okazać się nieistotne (ang. irrelevant) tzn. nie wpływają na proces predykcji. Inne cechy mogą być nadmiarowe (ang. redundant) tzn. nie dostarczają żadnych nowych informacji, a nawet mogą wprowadzać zakłócenia do projektowanego modelu [Ps04]. Dlatego przed przystąpieniem do zadania eksploracji przeprowa-

dzono redukcję cech wykorzystując teorię zbiorów przybliżonych. Do obliczeń zaproponowano algorytmy wykorzystujące działania na wektorach i macierzach w środowisku Matlab. Wyróżnione zostały następujące macierze:

- Macierz $A[n \times m]$ – macierz przedstawiająca system informacyjny, gdzie n ozn. liczbę obiektów, a m liczbę atrybutów;
- Macierz rozróżnialności $DMi[n \times m]$ – macierz będąca wynikiem porównań wartości atrybutów i -tego obiektu macierzy A ze wszystkimi pozostałymi obiektami. Jest ona odpowiednikiem jednej kolumny w macierzy rozróżnialności przedstawionej w tabeli 3. Kolumny odpowiadają porównywanym atrybutom, a wiersze wynikom porównania wartości kolumn przypisując:
 - 0 – jeżeli $i \geq j$, gdzie i, j są obiektami z macierzy A
 - 0 – jeżeli $i < j$ oraz atrybut nie rozróżnia i -tego i j -tego obiektu
 - 1 – jeżeli $i < j$ oraz atrybut rozróżnia i -ty i j -ty obiekt.
- Macierz zmiennych wejściowych tablicy prawdy $TT[p \times m]$, gdzie m jest liczbą atrybutów systemu informacyjnego, a $p=2^m$ jest liczbą możliwych kombinacji ustawień atrybutów.
- Wektor wartości funkcji tablicy prawdy $FA[p \times 1]$, gdzie $p=2^m$ jest liczbą możliwych kombinacji w tablicy prawdy, a m liczbą atrybutów systemu informacyjnego. Początkowo jest to wektor 1, w kolejnych porównaniach wartości wektora mogą przyjmować wartość 0 dla wierszy które się różnią.

Znalezienie zbioru reduktów polega na zdefiniowaniu tablicy prawdy dla zbioru opisanego macierzą A , a następnie na znalezieniu funkcji $f(A)$ w postaci kanonicznej. W rozwiązaniu zadania zaproponowano następujący algorytm:

Krok: Utworzenie macierzy zmiennych wejściowych tablicy prawdy TT dla m atrybutów

Krok: Utworzenie wektora wartości funkcji tablicy prawdy FA (ostateczna postać), $tempX_FA$ (postać po porównaniu j -tego wiersza macierzy rozróżnialności i wejść tablicy prawdy)

for $i=1:(n-1)$

% ilość obiektów jest mniejsza od 1 ponieważ ostatniego nie sprawdzamy

% obliczenia są przeprowadzane dla każdej "kolumny" macierzy rozróżnialności, czyli

% porównanie i -tego atrybutu z pozostałymi

Krok: utworzenie macierzy X o rozmiarach $n \times m$ która do wiersza i jest kopią macierzy A , a od wiersza i jest kopią wiersza i

Krok: różnica macierzy X i A , pola wskazujące atrybuty o tych samych wartościach będą miały po operacji wartość 0, pola wskazujące atrybuty o różnych wartościach będą miały po operacji wartość różną od 0; wynik -> $tempX$

Krok: utworzenie macierzy rozróżnialności poprzez zastosowanie funkcji $sign$ na macierzy uzyskanej w poprzednim kroku; wynik-> $tempX_DM$

Krok: Negacja wartości pól macierzy $tempX_DM$ w celu dostosowania jej do notacji 0-1, w której 0 określa atrybut wyróżniający wiersze, a 1 niewyróżniający; wynik -> $tempX_nDM$

Krok: Wyznaczenie liczby atrybutów wyróżniających sumując na podstawie $tempX_DM$; wynik-> $tempX_sum$

% z każdej policzonej kolumny (macierzy rozróżnialności) generowane jest wyjście funkcji rozróżnialności dla tablicy prawdy

```

for j= i+1:n
%zaczynamy od i+1 bo pomijamy wiersze zawierające 1 1 1
krok: porównanie każdego wiersza macierzy rozróżnialności tempX_nDM z każdym wierszem
tablicy decyzyjnej TT (suma logiczna or)
krok: zapisanie wyniku porównania w j-tym kroku do jFA; jeżeli liczba atrybutów wyróżniających
jest większa lub równa sumie dla j-tego wiersza macierzy tempX_sum to funkcja przyjmie wartość
1, w przeciwnym wypadku funkcja przyjmie wartość 0.
Krok: odświeżenie funkcji tempX_FA na podstawie jFA (część wspólna)
end
end
FA= tempX_FA; - ostateczna postać wektora wartości funkcji tablicy prawdy

```

Aby sprawdzić funkcjonalność algorytmu wybrano macierz zawierającą 100 rozróżnialnych obiektów. Do analizy wybrano losowy zbiór 10 cech. Wartości cech zostały wcześniej dyskretyzowane do 10 przedziałów równej długości. Tablica prawdy w tym zadaniu miała rozmiar $2^{10} \times 10$. Wartość wykładnicza określa ilość możliwych kombinacji wartości cech zbiorze $\{0,1\}$. Wartość funkcji wyjścia jest wektorem o 2^{10} elementach przyjmujących wartości $\{0,1\}$. W wyniku minimalizacji funkcji uzyskano tylko jeden redukt, który zawierał 8 atrybutów (cech).

Następnie algorytm został przetestowany na znacznie większym zbiorze zawierającym ok. 5 tysięcy obiektów. Do redukcji wybrano zestaw 12 słabo skorelowanych atrybutów. Minimalizacja funkcji rozróżnialności nie przyniosła jednak oczekiwanych rezultatów. Uzyskany redukt zawierał wszystkie atrybuty wejściowe.

Problemem z jakim należy zmierzyć się w tym algorytmie jest duża złożoność obliczeniowa. Przy tworzeniu macierzy rozróżnialności porównujemy ze sobą wszystkie obiekty. Złożoność obliczeniowa wyznaczenia macierzy rozróżnialności jest rzędu $O(n \cdot x^2)$, gdzie n – jest liczbą atrybutów, a x liczbą obiektów. Ponieważ operacje wykonywane są na macierzach to porównanie wartości atrybutów między dwoma obiektami odbywa się w jednym kroku. Złożoność można więc ograniczyć do $O(x^2)$. Utrudnienia powstają także przy tworzeniu tablicy decyzyjnej. Im większa liczba zredukowanego zbioru atrybutów tym więcej możliwych przypadków kombinacji, bo aż 2^n . Przy 10 atrybutach będziemy mieli 1024 możliwości, przy 12 atrybutach 4096, ale już przy 30 ponad 1 milion kombinacji. Należy przy tym pamiętać, że wyznaczenie wartości funkcji rozróżnialności polega na porównaniu każdego wiersza macierzy rozróżnialności z każdym wierszem tablicy prawdy co daje złożoność obliczeniową rzędu $O(2n \cdot x^2)$.

W postawionym zadaniu każdy obiekt jest opisany przez 217 atrybutów. Redukcja tak dużego zbioru byłaby bardzo czasochłonna. Dlatego metodę tą należałoby poprzedzić wstępną selekcją atrybutów opartą na przykład na korelacji omówioną we wcześniejszych pracach [StSz06, Szy06]. Przy współczynniku korelacji 0.6 zbiór atrybutów został zredukowany do 33. Dla takiej liczby ilość wierszy tablicy prawdy przekroczy 8,5 miliona kombinacji. Nadal będzie to zadanie o dużej złożoności. Rozwiązaniem problemu może okazać się zrównoleglenie zadania.

5. Podsumowanie

Ze względu na złożoność metody zbiorów przybliżonych w zastosowaniu do klasyfikacji komórek przeprowadzane będzie zrównoleglenie zadania. Opracowane algorytmy działają na macierzach i łatwo można tutaj wydzielić zadania równoległe. Wyselekcjonowane zbiory atrybutów zostaną wykorzystane do tworzenia modeli Data Mining. Planowane jest zastosowanie opracowanych algorytmów do analizy innych komórek nowotworowych.

Bibliografia

- [HaMa] , Hand D., Mannila H., P.Smyth, Eksploracja danych ISBN: 83-204-3053-4
- [Swi01] Świniarski R., Rough sets methods in feature reduction and classification, *Int. J. Appl. Math. Comput.*, 2001, Vol.11, No.3, 565-582
- [WaMa99] Walczak B., Massart D.L., "Rough sets theory", *Chemoetrics and Intelligent Laboratory Systems* 47, 1999, 1-16
- [Dom04] Dominik A. „Analiza danych z zastosowaniem teorii zbiorów przybliżonych”, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki, 2004
- [BCBK03] Borkowska E., Constantinou M., Binka-Kowalska A., Kałużewski B.: Diagnostyka raka pęcherza moczowego przy użyciu metody MSSCP (eksony 5-8 genu P53) i testu UroVysion, I Konferencja Użytkowników DNA Pointer System, Warszawa, 2003
- [Bro01] Brown T.A.: *Genomy*, Wydawnictwo Naukowe PWN, Warszawa 2001, ISBN 83-01-13439-9
- [PILo01] Plesch A., Loerch T.: Metafer – a Ultra Novel High Throughput Scanning System for Rare Cell Detection and Automatic Interphase FISH Scoring, Early Prenatal Diagnosis, Fetal Cells and DNA in the Mother, Present State and Perspectives, 12th Fetal Cell Workshop, Prague, May 2001, pp.329–339
- [Guz05] Guz T.: Poprawa efektywności klasyfikatora „Box Classifier” w systemie „Metafer”, XIII Konferencja „Sieci i Systemy Informatyczne”, Łódź, 2005.
- [Ps04] Piramuthu S.: “Evaluating feature selection methods for learning in data mining applications”; *European Journal of Operational Research* 156 (2004); p.483-494
- [StSz06] Stanisławski W., Szydłowska E.; „Analiza narzędzia Data Mining ORACLE 10g do klasyfikacji komórek nowotworowych w cytometrycznym systemie skaningowym” XII Konferencja użytkowników i deweloperów ORACLE, 17-20.10.2006 Zakopane – Kościelisko, str 251-263
- [Szy06] Szydłowska E. „Klasyfikacja komórek rakowych z wykorzystaniem technik eksploracji danych, Politechnika Opolska, Zeszyty Doktoranckie, 2006