

POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRONIKI I TECHNIK
INFORMACYJNYCH
INSTYTUT INFORMATYKI

Rok akademicki 2003/2004

PRACA DYPLOMOWA MAGISTERSKA

Paweł Terlecki

Wyznaczanie grup genów zależnych przy użyciu technologii mikromacierzy

Opiekun pracy:

dr hab. inż. Krzysztof Walczak

Ocena

.....

Podpis Przewodniczącego

Komisji Egzaminu Dyplomowego



Specjalność **Informatyka –
Inżynieria oprogramowania
i systemy informacyjne**

Data urodzenia **7 marzec 1980**

Data rozpoczęcia studiów **01 października 1999**

ŻYCIORYS

Urodziłem się 07.03.1980 r. w Warszawie. Po ukończeniu szkoły podstawowej, kontynuowałem naukę w XIV L.O. im. Stanisława Staszica w Warszawie. W szkole średniej uczęszczałem do klasy o profilu matematyczno-fizycznym z rozszerzonym programem informatyki. W październiku 1999 r. rozpocząłem studia na Wydziale Elektroniki i Technik Informatycznych Politechniki Warszawskiej. W październiku 2003 r. rozpocząłem studia na Wydziale Nauk Ekonomicznych Uniwersytetu Warszawskiego.

.....
podpis studenta

EGZAMIN DYPLOMOWY

Złożył egzamin dyplomowy w dniu2004 r.
z wynikiem
Ogólny wynik studiów

Dodatkowe wnioski i uwagi Komisji
.....
.....

STRESZCZENIE

Niniejsza praca ma na celu wprowadzenie w tematykę badania ekspresji genów za pomocą technologii mikrotablic. Objasnia podstawowe pojęcia charakterystyczne dla dziedziny oraz przedstawia ogólny sposób prowadzenia eksperymentów.

Najistotniejszą częścią pracy jest propozycja ścieżki analizy danych uzyskiwanych z doświadczeń. Obejmuje ona statystyczne badanie znaczenia genów, dyskretyzację wartości ciągłych, wyznaczanie reduktów oraz wzorców wyskakujących. Celem obliczeń jest porównanie wyników otrzymywanych przy stosowaniu różnych metod w zakresie oceny pojedynczych genów oraz znajdowania grup genów powiązanych. Jednocześnie przeprowadzone są rozważania formalne i eksperymentalne dotyczące pojęcia reduktu i wzorca wyskakującego.

Istotnym elementem pracy jest aplikacja Microarray Analyser. Pozwala ona na analizę danych zgodnie z opisaną ścieżką analizy. Implementacja obejmuje wszystkie przedstawione w pracy algorytmy.

Słowa kluczowe: ekspresja genów, mikromacierz, analiza wariancji, redukt, wzorzec wyłaniający

Finding of associated genes groups using microarray technology

ABSTRACT

The purpose of this thesis is to present the topic of gene expression analysis using microarray technology. It explains the principals of the domain and gives an outlook for and experiment conducting.

The most significant part contains a proposition of an analysis path for the data obtained in experiments. It comprises of statistical evaluation of gene importance, discretisation of continuous values, finding of reducts and emerging patterns. The main idea is to compare results of different methods in problems of gene evaluation and associated genes groups finding. The notions of reduct and jumping emerging pattern are taken under formal and experimental consideration.

The important part of this thesis is Microarray Analyser application. It allows conducting analysis according to the described path. All presented algorithms are covered by the implementation.

Key words: gene expression, microarray, analysis of variance, reduct, emerging pattern

Spis treści

1.	Wstęp	7
1.1.	Wprowadzenie	7
1.2.	Umiejscowienie w dziedzinie	8
1.3.	Cel pracy	9
1.4.	Zawartość merytoryczna	9
2.	Biologiczny zarys problemu	11
2.1.	Podstawy biochemiczne	11
2.1.1.	Białka	11
2.1.2.	Kwasy nukleinowe	11
2.1.3.	Kod genetyczny	13
2.1.4.	Ekspresja genów	13
2.1.5.	Biblioteki genowe	14
2.2.	Badanie ekspresji genów	15
2.3.	Dostępne technologie oparte o mikromacierze	16
2.4.	Schemat eksperymentu	16
2.5.	Pomiar poziomu ekspresji	19
2.6.	Normalizacja	20
2.7.	Rodzaje analizy danych	22
2.8.	Dane dostępne publicznie	22
2.9.	Przykład eksperymentu	23
2.10.	Standaryzacja opisu	24
2.11.	Koncepcja bazy danych mikromacierzy	25
2.12.	Specyfikacja MIAME	27
3.	Formalny opis danych	28
3.1.	Wprowadzenie	28
3.2.	Definicje	28
4.	Wyznaczanie zbioru genów znaczących	32
4.1.	Wprowadzenie	32
4.2.	Znaczenie	33
4.3.	Statystyczna ocena atrybutów	33
4.3.1.	Wprowadzenie	33
4.3.2.	Jednoczynnikowa analiza wariancji	34
4.3.3.	Statystyka χ^2	36
4.4.	Pojęcie zbioru znaczącego	38
4.4.1.	Ocena znaczenia zbioru atrybutów	38
4.4.1.1.	Przypadek atrybutów dowolnych	40
4.4.1.2.	Przypadek atrybutów ciągłych	40
4.4.1.3.	Przypadek atrybutów dyskretnych	40
4.4.2.	Ocena znaczenia zbioru atrybutów ze względu na podzbiory	41
4.4.3.	Wyznaczanie zbiorów znaczących	45
4.4.4.	Wyznaczanie zbioru λ ,1-sumacyjnie znaczącego	46
4.4.5.	Dobór parametrów zbiorów znaczących	47
5.	Dyskretyzacja	48
5.1.	Wprowadzenie	48
5.2.	Definicje	48
5.3.	Podstawowe metody	49
5.3.1.	Równa szerokość przedziałów	50
5.3.2.	Równa głębokość przedziałów	51

5.3.3.	Podjęcie zstępujące.....	52
6.	Wyznaczanie rozróżniających grup genów.....	55
6.1.	Wprowadzenie.....	55
6.2.	Elementy teorii zbiorów przybliżonych.....	55
6.3.	Realizacja wybranych pojęć.....	64
6.3.1.	Struktury rozróżnialności obiektów.....	64
6.3.2.	Funkcja rozróżnialności.....	64
6.4.	Metody znajdowania wszystkich reduktów.....	65
6.4.1.	Podjęcie dokładne.....	65
6.4.2.	Podjęcie zachłanne.....	66
6.4.3.	Podjęcie heurystyczne.....	67
6.4.4.	Dekompozycja przestrzeni.....	67
6.5.	Metody wyszukiwania minimalnych k-zbiorów trafieniowych.....	68
6.5.1.	Wprowadzenie.....	68
6.5.2.	Podjęcie dokładne.....	68
6.5.3.	Podjęcie losowe.....	68
6.5.4.	Podjęcie ewolucyjne.....	68
7.	Wyznaczanie charakterystycznych wzorców ekspresji.....	72
7.1.	Wprowadzenie.....	72
7.2.	Znaczenie dla badania ekspresji.....	73
7.3.	Definicje i twierdzenia.....	73
7.4.	Wyznaczanie wzorców wyskakujących.....	83
7.4.1.	Wprowadzenie.....	83
7.4.2.	Przestrzeń wsparcia wzorców.....	83
7.4.3.	Podstawowe metody wyznaczania wzorców wyłaniających.....	84
7.4.4.	Wyznaczanie płotów horyzontalnych.....	85
7.4.5.	Wyznaczanie różnicy płotów.....	86
7.4.6.	Wyznaczanie przestrzeni wzorców wyskakujących.....	87
7.4.7.	Analiza efektywności.....	90
8.	Zależność reduktów i wzorców wyskakujących.....	92
8.1.	Wprowadzenie.....	92
8.2.	Rozważania formalne.....	92
8.3.	Prowadzenie analizy wyników.....	96
8.3.1.	Wprowadzenie.....	96
8.3.2.	Przypadek reduktów.....	97
8.3.3.	Przypadek wzorców wyskakujących.....	98
8.3.4.	Porównanie miar istotności genów.....	100
8.4.	Analiza postaci zbiorów wynikowych.....	101
9.	Wyniki eksperymentalne.....	104
9.1.	Sposób prowadzenia analizy.....	104
9.2.	Wyniki eksperymentów.....	104
9.2.1.	Zbiór danych „Ontario”.....	105
9.2.2.	Zbiór danych „Princeton”.....	112
9.3.	Omówienie wyników.....	119
9.3.1.	Ocena pojedynczych genów.....	119
9.3.2.	Grupy genów współzależnych.....	120
9.3.3.	Zbiór reduktów i przestrzeń wzorców wyskakujących.....	122
10.	Opis programu testowego Microarray Analyser.....	124
10.1.	Informacje ogólne.....	124
10.2.	Format danych wejściowych.....	124

10.3. Prowadzenie analizy zbioru danych.....	127
10.4. Propozycje rozwoju	134
11. Podsumowanie.....	136
12. Materiały	139
12.1. Bibliografia.....	139
12.2. Zasoby internetowe	141

1. Wstęp

1.1. Wprowadzenie

W ostatniej dekadzie obserwujemy szczególnie silny rozwój bioinformatyki. Jest ona dziedziną interdyscyplinarną, która wykorzystuje współczesne metody analizy danych do zagadnień biologicznych. Dodatkowy rozpęd nadaje badaniom postęp w zakresie inżynierii genetycznej, której zdobycze wykorzystywane są także w genetyce, biochemii i biologii molekularnej.

Hipoteza postawiona przez Watsona i Cricka, wskazująca łańcuch kwasu nukleinowego zawarty w komórkach organizmu jako nośnik informacji dziedzicznej, wprowadziła genetykę na nowe tory.

Jednym z pierwszych zadań, które postawiła przed sobą bioinformatyka, było poznanie struktury i funkcjonalności elementów łańcucha DNA różnych organizmów. Cel zaowocował wykształceniem zaawansowanych metod identyfikacji genów oraz zasad klasyfikacji gromadzonych informacji. Dla wielu gatunków istnieją obecnie kompletne biblioteki opisujące poszczególne elementy ich DNA.

Z tematyką silnie powiązana jest analiza modyfikacji informacji dziedzicznej zachodząca pomiędzy pokoleniami. Do analizy wykorzystywane są różne metody uliniawiania sekwencji oraz drzewa filogenetyczne pokazujące kolejne przemiany DNA osobników posiadających wspólnego przodka.

Identyfikacja genów poszczególnych gatunków otworzyła drogę do analizy funkcji pełnionych przez nie w komórkach. Geny determinują syntezę białek, które odpowiadają za działanie całego organizmu. Ze względu na złożoność procesów życiowych wyznaczanie powiązań pomiędzy aktywnością określonych grup genów a obserwowalnym stanem organizmu odbywa się najczęściej z pominięciem etapów pośrednich. Aktywność określonych genów kojarzona jest bezpośrednio z konkretnymi stanami komórek. Wiedza ta pozwala na znajdowanie grup funkcjonalnych genów odpowiedzialnych za rozważane zjawiska. Wśród wielu zastosowań praktycznych należy wymienić poznawanie podłoża chorób genetycznych, testowanie działania leków oraz toksykologię. Obecnie najbogatszej wiedzy na temat roli genów dostarczają eksperymenty, które wykorzystują technologie mikromacierzy.

Kolejny etap badań dotyczy budowy i funkcji białek syntetyzowanych w komórkach organizmu. Zagadnienie to jest o wiele bardziej złożone od analizy kwasów nukleinowych, gdyż przestrzenna struktura białka ma kluczowe znaczenie dla jego właściwości i pełnionych przezeń funkcji. Zdobywana wiedza jest szczególnie istotna w dziedzinach związanych z odbudową określonych organów oraz wytwarzaniem materiałów organicznych o zadanych właściwościach.

Postęp w prężnie rozwijającej się dziedzinie w znaczący sposób wpływa na kształt obecnego świata. Jednym z najbardziej spektakularnych sukcesów jest ukończenie w 2003 roku projektu dotyczącego poznania genomu ludzkiego (Human Genome Project). Projekt został ukończony w ciągu 13 lat, w czasie o dwa lata krótszym niż pierwotnie założony. Obecnie znamy kompletną mapę genów człowieka. Następne wyzwanie dotyczy odkrycia znaczenia poszczególnych genów w obserwowanych procesach życiowych.

Pomimo dużej intensywności badań na całym świecie bioinformatyka kryje wiele nierozwiązanych dotąd zagadnień. Najbliższe lata z całą pewnością otworzą przed nami nowe horyzonty możliwości.

1.2. Umiejscowienie w dziedzinie

Praca niniejsza zajmuje się analizą danych opisujących ekspresję genów. Badanie ekspresji odbywa się obecnie za pomocą trzech podstawowych technik: szeregowej analizy ekspresji genów, mikromacierzy oligonukleotydowych i mikromacierzy cDNA. W ostatnich latach szczególną popularność zdobyły dwie ostatnie spośród wymienionych metod. Ze względu na ogólne podobieństwo obu podejść zakładam w pracy, że punktem wyjścia do dalszego przetwarzania jest eksperyment przeprowadzony zgodnie z dowolną z nich.

Ogromna ilość danych pochodzących z doświadczeń zaowocowała dużymi możliwościami pozyskiwania wiedzy na temat genów. Zgodnie ze sposobem prowadzenia analizy rozpatrywane zagadnienia można podzielić na kilka podstawowych grup.

Technologia mikromacierzy jest powszechnie używana do identyfikacji genów [2]. Sekwencje nukleinowe rozważanego organizmu są rozpoznawane poprzez hybrydyzację ze znanymi genami organizmów prostszych. Podejście to nie jest nakierunkowane na badanie intensywności ekspresji.

W zależności od sposobu przedstawienia danych opisujących wielkość ekspresji otrzymujemy dwie różne rodziny zagadnień. W obu przypadkach możemy stosować metody uczenia bez nadzoru i pod nadzorem [9].

W pierwszym przypadku z rozpatrywanymi genami związane są wektory poziomów ekspresji uzyskanych w kolejnych eksperymentach [9,40]. Podejście to skupia się na wykrywaniu genów o podobnych wzorcach aktywności. W ten sposób identyfikowane są grupy genów ulegających koekspresji w określonych warunkach.

Do analizy bez nadzoru wykorzystywane jest najczęściej grupowanie hierarchiczne, algorytm k-s rednich, samoorganizujące się mapy (SOM – Self-organizing maps) i metoda składowych głównych (PCA – Principal Component Analysis). Uczenie pod nadzorem wykorzystuje posiadaną wiedzę biologiczną do nakierowywania algorytmów grupujących. Szczególne znaczenie dla znajdowania powiązań między genami odgrywa klasyfikacja oparta o informacje na temat grup genów o zbieżnej ekspresji. Spotykane w literaturze propozycje bazują na różnych popularnych klasyfikatorach, przy czym najczęściej wykorzystywanym rozwiązaniem jest maszyna wektorów nośnych (SVM – Support Vector Machine).

Zupełnie innego spojrzenia dostarcza potraktowanie każdej hybrydyzacji jako punktu w przestrzeni, którego współrzędne odpowiadają poziomom ekspresji kolejnych genów. Duże rozmiary genomów decydują o wysokiej wymiarowości przestrzeni. Z tego względu analiza wyników wymaga stosowania wyłącznie algorytmów o dobrej skalowalności.

W pierwszej kolejności podejście jest nastawione na znajdowanie powiązań pomiędzy hybrydyzacjami, które odpowiadają różnym stanom badanych komórek. Najczęstszym zastosowaniem praktycznym jest budowa klasyfikatorów wykorzystujących pomiar ekspresji dla ustalenia czy badana tkanka posiada określone cechy. Przykładowo, może ono dotyczyć wykrywania nowotworów [19], skutków zastosowanego leczenia lub skażenia toksynami.

Za pomocą metod statystycznych możliwe jest też wyznaczenie pojedynczych genów, które znacząco zmieniają swoją ekspresję dla różnych stanów organizmów. W ten sposób otrzymujemy informacje o genach, które mogą być związane z badanym zjawiskiem.

Głębsza analiza pozwala na wyznaczenie powiązań pomiędzy określonymi wzorcami ekspresji oraz odpowiadającymi im cechami. Wzorce nie tylko mogą być wykorzystywane do klasyfikacji. Stanowią one również podstawę do wyznaczania zależności międzygenowych bardziej złożonych niż koekspresja. Metody analizy wykorzystują między innymi ideę wzorców wyłaniających [31] oraz znane klasyfikatory, w tym oparte o teorię zbiorów przybliżonych [21].

W pracy zajmuję się analizą danych pochodzącą z ostatniej z prezentowanych tu klas zagadnień. Skupiam się na identyfikacji istotnych genów oraz grup genów, które są silnie związane z danym zjawiskiem.

1.3. Cel pracy

Praca ma podstawowe zadanie wprowadzenia czytelnika w tematykę badania ekspresji genów. Porusza elementarne zagadnienia dotyczące prowadzenia eksperymentów i analizy ich rezultatów. Pokazuje obecne zaawansowanie dziedziny oraz cele, jakie w jej zakresie stawia przed sobą bioinformatyka.

Rozpatrywana jest klasa eksperymentów, które wykorzystują do badania ekspresji genów technologię mikromacierzy. W szczególności skupiam się na doświadczeniach dających punkt wyjścia do uczenia pod nadzorem.

W pracy zaprezentowana została ścieżka analizy danych związanych z ekspresją, której celem jest określenie grup genów zależnych odpowiedzialnych za regulację określonych procesów. Obejmuje ona kolejno: wstępny wybór genów znaczących za pomocą metod statystycznych, dyskretyzację wartości ciągłych oraz wyznaczanie relacji pomiędzy grupami genów i określoną w eksperymencie klasą decyzyjną. Do badania grup genów wykorzystane zostały niezależnie podejścia związane z teorią zbiorów przybliżonych i pojęciem wzorców wyłaniających.

Ponadto dane na temat ekspresji zostały wykorzystane jako tło do porównania postaci i własności zbioru reduktów względnych otrzymanych dla pewnej tablicy decyzyjnej z przestrzenią wzorców wyskakujących wyznaczonych dla rodziny baz danych odpowiadających tej tablicy. W szczególności na podstawie obu pojęć zaproponowano miary istotności pojedynczych genów, które zostały następnie porównane z oszacowaniami statystycznymi.

1.4. Zawartość merytoryczna

Rozdział 2 daje ogólne rozeznanie w biologicznych podstawach analizy ekspresji genów. Dużo miejsca poświęcone jest technologii mikromacierzy oraz sposobowi prowadzenia i dokumentowania doświadczeń.

Rozdział 3 precyzuje formalny opis danych otrzymywanych w wyniku eksperymentów. Stanowi on punkt wyjścia do kolejnych etapów analizy danych. W rozdziale tym wprowadzony jest również główny przykład, który będzie śledzony w całej pracy.

Rozdział 4 dotyczy zagadnienia znajdowania zbioru genów znaczących. Zawiera propozycję ogólnego podejścia do problemu wyłonienia zestawu atrybutów istotnych dla dalszej analizy oraz omawia powszechnie stosowane w praktyce metody statystyczne dla atrybutów ciągłych i dyskretnych. Ograniczenie zbioru genów jest ważnym etapem poprzedzającym zastosowanie metod bardziej zaawansowanych, gdyż są one zwykle mało efektywne dla problemów o dużej wymiarowości.

Dane dotyczące ekspresji genów mają charakter ciągły. Rozdział 5 szczegółowo omawia popularne algorytmy dyskretyzacji. Ich zastosowanie umożliwia analizę z wykorzystaniem metod odkrywania wiedzy.

W rozdziale 6 rozważane jest zagadnienie znajdowania grup genów o dużej zdolności rozróżniania obiektów. Do tego celu wykorzystane zostały elementy teorii zbiorów przybliżonych. Dużo miejsca poświęcone jest wyznaczaniu reduktów względnych tablicy decyzyjnej. Dla tego problemu przytoczone zostały podejścia spotykane w literaturze, a także zaproponowany algorytm przybliżony oparty o dekompozycję przestrzeni możliwych podzbiorów atrybutów.

Rozdział 7 opisuje ideę wzorców wyłaniających. Szczegółowo omówione jest wyznaczanie przestrzeni wzorców wyskakujących i jej zastosowanie do wyznaczania charakterystycznych profili ekspresji genów.

W rozdziale 8 zajmujemy się porównaniem pojęcia reduktu i wzorca wyskakującego. Przedstawione są podstawowe przesłanki teoretyczne przemawiające za powiązaniem tych pojęć oraz eksperymentalne metody służące odniesieniu właściwości zbioru reduktów do przestrzeni wzorców wyskakujących. W tym celu zdefiniowane zostały odpowiednie funkcje oceny istotności atrybutów.

Rozdział 9 zawiera opis analizy danych oraz jej rezultaty dla dwóch rzeczywistych zbiorów danych uzyskanych za pomocą techniki mikromacierzy. Dla badanych zestawów przeprowadzono porównanie wyznaczonych reduktów i wzorców wyskakujących za pomocą metod zaproponowanych w rozdziale 8. Otrzymane wyniki empiryczne zostały podsumowane omówieniem i odniesieniem do analiz prowadzonych przez inne zespoły.

Rozdział 10 obejmuje ogólny opis aplikacji wykorzystywanej do wykonania testów. Omówiony został przede wszystkim format plików wejściowych oraz sposób prowadzenia analizy.

Rozdział 11 zawiera podsumowanie całej pracy. Podkreślone zostały podstawowe elementy pracy, osobisty wkład oraz płynące z badań wnioski.

Rozdział 12 zawiera wykaz publikacji i zasobów internetowych wykorzystanych w pracy.

2. Biologiczny zarys problemu

2.1. Podstawy biochemiczne

2.1.1. Białka

Fundamentalne znaczenie dla funkcjonowania organizmów żywych mają białka. Są one złożonymi związkami chemicznymi o charakterze polimerycznym. Na łańcuch związku składają się aminokwasy połączone wiązaniami peptydowymi. Z tego względu białko możemy postrzegać również jako polipeptyd.

Aminokwasy są prostymi związkami organicznymi. Na pojedynczy polipeptyd przypada od 100 do 1500 aminokwasów. W białkach biogennych występują one dokładnie w 20 różnych rodzajach [44].

Kolejność występowania aminokwasów w łańcuchu określana jest mianem struktury pierwszorzędowej. Definicje wyższych struktur białkowych związane są z przestrzenną budową łańcucha oraz zależnościami pomiędzy oddzielnymi łańcuchami.



Rys. 2.1.1-1. Przykładowa struktura przestrzenna białka

Białka pełnią wielorakie funkcje w organizmie. Pod postacią enzymów katalizują niemal wszystkie reakcje biochemiczne przyspieszając czas ich zajścia o kilka rzędów wielkości. Jednocześnie w wielu złożonych procesach odgrywają rolę regulacyjną. Jako przeciwciała tworzą podstawową barierę odpornościową dla bakterii, wirusów i innych substancji obcych. Z kolei jako składnik błon komórek nerwowych są naturalnymi receptorami sygnalizującymi stan środowiska zewnętrznego. Transportują oraz magazynują substancje organiczne i nieorganiczne poprzez tworzenie z nimi związków tymczasowych. Dla skóry, kości i tkanki łącznej białka są głównym elementem budulcowym. Jeśli w organizmie brakuje innych źródeł energii zużywane są one również w wysokoenergetycznych procesach spalania.

Ze względu na swoją wszechobecność białka stanowią podstawową dziedzinę zainteresowań badaczy. Szczególny nacisk kładziony jest na analizę funkcji sterujących, gdyż właśnie one implikują zachowanie organizmu w określonych warunkach.

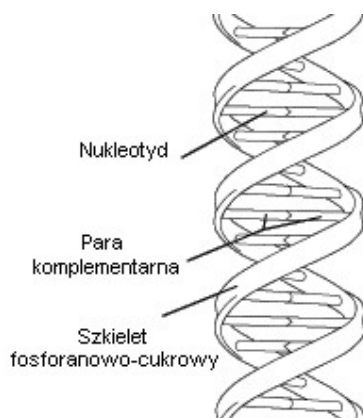
2.1.2. Kwasy nukleinowe

Informacja o wszelkich białkach produkowanych przez organizm zapisana jest w łańcuchu kwasu nukleinowego. W zależności od stopnia skomplikowania organizmu nośnik stanowi inny kwas.

W prostych organizmach jednokomórkowych (prokaryota) powyższą funkcję pełni kwas rybonukleinowy RNA. Jego cząsteczka często stanowi jądro komórki. Ma on strukturę polimeryczną, przy czym pojedynczym monomerem jest nukleotyd. Każdy nukleotyd składa się z 5-węglowego cukru (rybozy), reszty fosforanowej i zasady azotowej. Ze względu na różne typy zasad azotowych występują cztery rodzaje nukleotydów: adenina (A), guanina (G), cytozyna (C) i uracyl (U). Dwie pierwsze określa się mianem puryn, a pozostałe pirymidyn. Cząsteczki RNA odgrywają także istotną rolę przy dekodowaniu informacji genetycznej zarówno u organizmów prostych, jak i bardziej złożonych.

U organizmów wyższych (eukaryota) informacja o białkach zakodowana jest w kwasie dezoksyrybonukleinowym DNA. Podobnie, jak RNA jest on polimerem nukleotydów, przy czym uracyl zastępuje inna pirymidyna – tymina (T). Dla budowy i replikacji kwasu istotną cechą nukleotydów jest ich komplementarność. Terminem tym określa się możliwość powstawania silnych wiązań wodorowych w konfiguracjach: adenina i tymina (również uracyl) oraz guanina i cytozyna.

W stanie naturalnym kwas DNA występuje w dwóch komplementarnych niciach połączonych mostkami wodorowymi i przyjmuje w przestrzeni kształt podwójnej helisy. Do odtworzenia pełnej informacji wystarczy jedna z nich.



Rys. 2.1.2-1. Schemat podwójnej helisy kwasu DNA

Organizmy wyższe składają się z komórek o dobrze wyodrębnionym jądrze komórkowym. Ze względu na długość nici kwas DNA jest mocno poskręcany i upakowany do postaci struktur takich jak chromatyna lub chromosomy. Struktury te znajdują się w jądrach komórkowych jądrzastych komórek organizmu. Bez względu na własne przeznaczenie, każda z tych komórek posiada pełną informację o białkach syntetyzowanych w całym organizmie.

Struktura przestrzenna kwasów, podobnie jak białek, jest skomplikowana. Obserwuje się między innymi różne stopnie skręcenia cząsteczki oraz występowanie dodatkowych słabych wiązań pomiędzy przylegającymi regionami nici. Najistotniejsza dla syntezy białek jest pierwszorzędowa struktura kwasu, która określa liniową kolejność nukleotydów w łańcuchu.

Z rozmnażaniem się komórek nierozdzielnie związany jest proces replikacji DNA. Zachodzi on tuż przed podziałem jądra komórkowego na jądra potomne i zapewnia identyczność przekazywanej im informacji genetycznej. Replikacja polega na podwajaniu się cząsteczki kwasu dezoksyrybonukleinowego. Jego pierwszym etapem jest rozwijanie i rozdzielanie komplementarnych łańcuchów polinukleotydu. Następnie każdy pojedynczy łańcuch staje się

matrycą do budowy nowego łańcucha komplementarnego. W ten sposób z pierwotnej podwójnej helisy DNA powstają dwie nowe, przy czym każda zawiera po jednym łańcuchu oryginalnym i po jednym nowym.

Ponieważ dla dalszych rozważań nie jest istotne, który z kwasów jest nośnikiem informacji, więc ze względu na przejrzystość dalsze rozważania będą prowadzone dla kwasu DNA. Jeśli nie zostanie wyraźnie zaznaczona różnica, to *implicite* należy przyjąć, że odnoszą się one również do organizmów prostszych. Szersze omówienie struktury kwasów nukleinowych można znaleźć w pracy [44].

2.1.3. Kod genetyczny

Badania nad molekularną strukturą kwasów nukleinowych rozpoczął głośny artykuł Watsona i Cricka pod tytułem „Molecular Structure of Nucleic Acids”, który ukazał się w „Nature” w kwietniu 1953 roku. Dzięki tej pracy genetyka zyskała dla dalszych badań bogate zaplecze biochemiczne pozwalające na weryfikację i formułowanie nowych hipotez. Od tego momentu powszechnie przyjmuje się, że cała informacja dziedziczna organizmu zapisana jest w cząsteczce DNA.

Za podstawową jednostkę informacji dziedzicznej uważany jest gen. Pierwotnie pojęcie to miało charakter abstrakcyjny, obecnie odnosi się ono do określonego odcinka kwasu nukleinowego. Miejsce w ramach sekwencji, gdzie znajduje się gen nazywane jest *locus*.

Geny nie są ciągłymi odcinkami DNA. Odcinki kwasu nie kodujące żadnego elementu końcowego produktu genu są nazywane intronami. Odpowiednio pozostałe odcinki składające się na informację niesioną przez geny nazywamy eksonami [44].

Zespół genów zawarty w pojedynczej nici kwasu nukleinowego organizmu określamy mianem genomu. Zauważmy, że organizm eukariotyczny posiada dwa genomy, gdyż jego materiał genetyczny zawarty jest w podwójnej helisie DNA.

Zbiór wszystkich genów organizmu nazywany jest genotypem. Ponieważ pojedynczy gen koduje jedno określone białko, więc cały genotyp zawiera informacje o wszystkich białkach, które mogą być syntetyzowane w organizmie. Zbiór cech organizmu określa się mianem fenotypu. Przyjmuje się, że genotyp implikuje fenotyp.

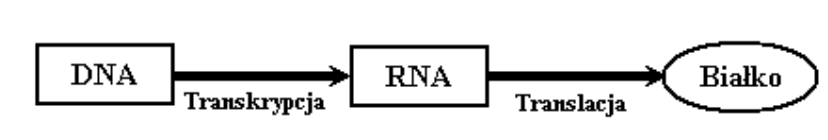
2.1.4. Ekspresja genów

Ekspresja genu jest procesem wykorzystywania informacji zawartej w genie do produkcji określonego białka. Składają się na nią: transkrypcja genowa sekwencji DNA lub RNA oraz, dla większości genów, translacja sekwencji RNA do białka. Pod pojęcie to często zalicza się składanie RNA oraz transport produktów genów do odpowiednich miejsc w komórce. Z punktu widzenia otrzymywanych produktów proces określa się również mianem biosyntezy białek.

Składanie RNA (*splicing*) obejmuje proces usuwania intronów z cząsteczki RNA i łączenia eksonów tak, by powstał funkcjonalny RNA. Jest to etap przygotowujący łańcuch do transkrypcji [43].

Transkrypcja jest syntezą jednoniciowego RNA na matrycy DNA. Proces zachodzi przy obecności odpowiedniego dla danego genu enzymu polimerazy RNA. Odpowiada ona za odnalezienie początku genu, rozplecenie helisy DNA i przeprowadzenie samego przepisania. Transkrypcja informacji genetycznej umożliwia wyniesienie jej z obszaru jądra do cytoplazmy komórki [43,44].

Następnie kod genetyczny otrzymanego RNA jest przetłumaczany na cząsteczkę białka. Etap ten nazywany jest translacją i zachodzi w specjalnych strukturach komórki zwanych rybosomami.



Rys. 2.1.4-1. Schemat biosyntezy białek

Wytwarzanie produktów genów przez organizm określane jest regulacją aktywności genów. U bakterii regulacja sprowadza się do produkcji tylko tych białek, które są w danej sytuacji niezbędne. W tym wypadku śledzenie zmiany ilości produktu może być potencjalnie oszacowane przez zmianę tempa transkrypcji, okresu półtrwania matrycowej postaci RNA i zmianę wydajności translacji. W przypadku organizmów eukariotycznych komórki są specjalizowane i średnio tylko 15% genomu ulega ekspresji. Zestaw genów aktywnych określa cechy i rolę komórki. Regulacja ekspresji odbywa się głównie poprzez zmianę szybkości transkrypcji. Szersze omówienie można znaleźć w pracy [48].

Reakcję komórki na daną sytuację można analizować poprzez śledzenie procesu transkrypcji. Z kolei intensywność zachodzenia tego etapu łatwo sprowadzić do pomiaru stężenia RNA w cytoplazmie. Mówimy wtedy o badaniu poziomu ekspresji genu. Podejście takie pozwala na analizę zachowania organizmu w danej sytuacji.

2.1.5. Biblioteki genowe

Dla systematycznego badania organizmów niezbędne okazało się zbudowanie bibliotek zawierających ich pełen materiał genetyczny. W przypadku większości organizmów nie jest możliwe bezpośrednie ekstraktowanie genomu ze względu na brak wystarczającej ilości materiału badawczego. Z tego względu istotną rolę w tworzeniu repozytoriów odgrywa proces klonowania sekwencji DNA.

Klonowanie DNA ułatwia manipulację fragmentami genomu dzięki ich niezależnemu namnażaniu. W fazie początkowej łańcuch DNA jest wyprowadzany z organizmu dawcy, a następnie jest rozdzielany na izolowane fragmenty za pomocą technik elektroforetycznych. Wybrany fragment wprowadzany jest do organizmu biorcy, najczęściej bakterii, za pomocą wektorów – niewielkich cząsteczek DNA zdolnych do autonomicznej replikacji w środowisku komórki biorcy lub integracji z jego materiałem genetycznym. Właściwe namnażanie zachodzi w organizmie biorcy. Szerzej w pracy [44].

Bibliotekę genową sporządza się dla wybranego gatunku organizmu i jest ona zbiorem sekwencji jego sklonowanego DNA. Istotne jest oszacowanie, czy dana biblioteka faktycznie opisuje cały materiał genetyczny organizmu, cechę tę określa się mianem reprezentatywności. Jej zapewnienie związane jest ze sporządzeniem określonej liczby klonów, tak aby prawdopodobieństwo zapisania całego genomu było odpowiednio wysokie.

Wyróżniamy dwa zasadnicze rodzaje bibliotek genowych. Pierwszy z nich to biblioteka genomowa, która budowana jest bezpośrednio na genomowym DNA organizmu. Łańcuch jest izolowany z organizmu i losowo rozcinany na odcinki o wielkości odpowiedniej dla klonowania.

Drugi typ stanowi biblioteka cDNA. Jest ona populacją specyficznych, jednoniciowych cząsteczek kwasu DNA, określanych mianem cDNA. Cząsteczki te uzyskuje się poprzez odwrotną transkrypcję kwasu DNA na matrycowym RNA wytwarzanym w komórkach danego organizmu. Łańcuchy RNA można prosto wyizolować ze środowiska komórki poprzez wykorzystanie ich własności magnetycznych. Istotną cechą stosowanego podejścia jest nieobecność w uzyskanym produkcie tych sekwencji łańcucha, które zostały wycięte na etapie składania. W ten sposób cDNA zawiera wyłącznie materiał efektywny z punktu widzenia syntezy białek. Zauważmy dodatkowo, że badając określone narządy lub tkanki rozpatrywanego organizmu uzyskujemy bardziej wyrazisty obraz tej części genomu, który ulega w nich najsilniejszej ekspresji. W ten sposób dla wybranej grupy komórek otrzymujemy większe stężenie kopii zestawu genów, które są związane z jej funkcjonowaniem w aktualnym stanie. Fakt ten dostarcza wiedzy na temat powiązania określonych genów z rolą pełnioną przez komórke, ale jednocześnie utrudnia stworzenie reprezentatywnej biblioteki.

Dla organizmów prokariotycznych na ogół buduje się biblioteki genomowe. Uzasadnione to jest zarówno stosunkowo małym rozmiarem całości materiału genetycznego, jak i dużą niestabilnością matrycowego RNA, która utrudnia otrzymywanie cDNA. W przypadku organizmów eukariotycznych dużo bardziej użyteczne jest sporządzanie bibliotek cDNA, gdyż zawierają one wyłącznie sekwencje transkrybowane, a więc opisują profil syntetyzowanych w komórce białek. Z tego względu biblioteki te są powszechnie wykorzystywane w badaniach nad ekspresją genów.

2.2. Badanie ekspresji genów

Pomiar wielkości ekspresji jest podstawowym źródłem informacji na temat znaczenia oraz funkcji genów. Informacja na temat wysokości poziomów ekspresji może być wykorzystywana wielorako. Po pierwsze stanowi podstawę do opisu zachowania komórki wobec danych warunków środowiskowych lub zastosowanego leczenia, co pozwala na konstrukcję skutecznych klasyfikatorów. Z drugiej strony pomiar jest dobrym punktem wyjścia do bezpośredniego odkrywania wiedzy na temat funkcjonalności poszczególnych genów.

Obok rozszerzenia zakresu informacji dotyczących działania organizmów żywych, badania mają znakomity zakres zastosowań praktycznych. Są one szczególnie cenne dla medycyny pozwalając na identyfikację przyczyn chorób (również o podłożu genetycznym), a także na dokładne testowanie wpływu proponowanych leków na organizmy chore. W przemyśle często wykorzystywane są do wykrywania toksyn na podstawie reakcji mikroorganizmów na dane środowisko.

Poprawny pomiar ekspresji powinien jednocześnie dawać odpowiedź na temat istotnego dla danej sytuacji zakresu genomu. Wczesne metody, między innymi szeregową analizę ekspresji genów (Serial Analysis of Gene Expression, SEGA), mogły śledzić wyłącznie zachowanie pojedynczych genów. Okazują się one nieprzydatne w badaniu złożonych zależności międzygenowych. Wymóg jednoczesnego pomiaru dla dużej liczby genów spełniła dopiero technologia oparta o analizę mikromacierzy (GEMA – Gene Expression Microarray Analysis). W pracy badanie ekspresji genów jest utożsamiane z GEMA, gdyż jest to obecnie jedyne efektywnie stosowane podejście. Mikromacierze znajdują również zastosowanie w identyfikacji sekwencji genetycznych, np. wyodrębnianiu genów.

Technologia mikromacierzy pozwala monitorować transkrypcję genów w danej sytuacji i badać jak aktywność określonych genów objawia się w fenotypie. Dla pojedynczego pomiaru wynikiem jest wektor wartości, które odpowiadają genomowi organizmu. W zależności od celu

eksperymentu pomiary są powtarzane dla organizmów w określonych stanach i środowiskach. Doświadczenie może dotyczyć badania kolejnych stadiów cyklu życia komórki, porównania tkanek o różnej charakterystyce, np. posiadanych schorzeniach, chemicznym skażeniu.

Niestety już samo podejście pociąga za sobą istotne problemy. Po pierwsze przygotowanie eksperymentu jest skomplikowane i kosztowne. Po drugie, pomimo ciągłego rozwoju rozwiązań technicznych, otrzymywane dane są silnie zaszumione.

Dodatkowo, sama dziedzina kryje w sobie o wiele poważniejsze trudności natury nietechnicznej. Przede wszystkim istnieje bardzo wiele czynników, również niegenetycznych, mających wpływ na obserwowane zachowanie organizmu. W konsekwencji eksperymenty związane są z dużą liczbą danych wynikowych. Obecnie tylko niektóre dostępne algorytmy mogą być stosowane do efektywnej analizy tych rezultatów. Ponadto, pomiędzy badanymi czynnikami istnieje złożona sieć zależności, co skutecznie utrudnia izolowane badanie i opis ich faktycznego znaczenia. Współdziałające geny należy traktować, jako jednolite grupy funkcjonalne. Przy czym grupy te często nie są rozłączne.

2.3. Dostępne technologie oparte o mikromacierze

Mikromacierz jest prostokątną płytką wykonaną ze szkła lub z substratów nylonowych. Jest ona podzielona na komórki, w których umieszczane są odcinki kwasu nukleinowego. W zależności od rozmiaru komórek czasem stosuje się dwa oddzielne pojęcia: makromacierze (przekątna komórki ok. 300 mikronów) oraz mikromacierze (przekątna poniżej 200 mikronów). Mniejszy rozmiar komórki utrudnia nanoszenie kwasu, ale umożliwia umieszczanie większego zakresu genomu na macierzy.

Pomimo wspólnej zasady działania i postaci otrzymywanych danych wyjściowych istnieją różne technologie związane z eksperymentami dotyczącymi mikromacierzy. Definiują one zarówno budowę płytki, sposób umieszczenia na niej kwasu, jak i odczytu informacji końcowej.

W praktycznych zastosowaniach dominują dwie metody. Pierwsza została zaproponowana na uniwersytecie w Stanford. Wykorzystuje ona odpowiednio oznaczone sekwencje cDNA, które mogą być nałożone bezpośrednio na powierzchnie płytki. Nanoszenie materiału może odbywać się ręcznie lub automatycznie w zależności od rozmiaru komórek płytki. Do pomiaru jest stosowany laser dwukanałowy. Ogólnie technika zapewnia elastyczność w projektowaniu eksperymentów i jest relatywnie tania.

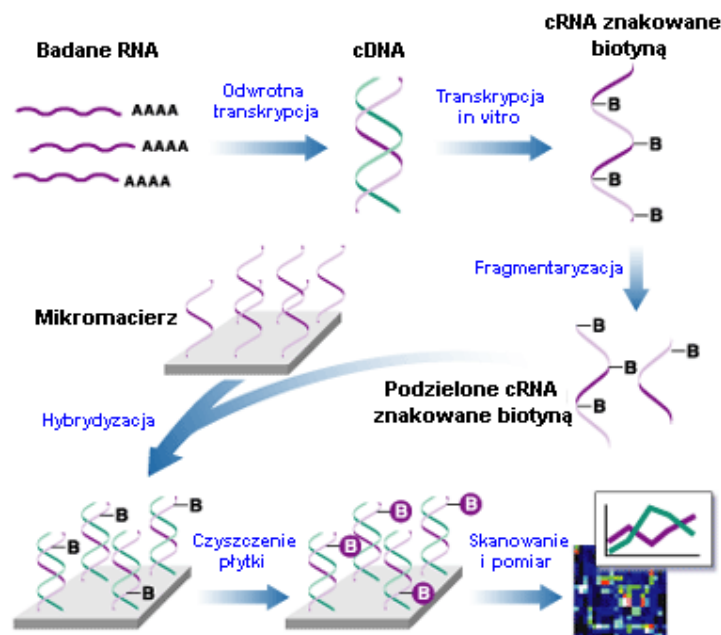
Drugie rozwiązanie oryginalnie zostało zaproponowane przez firmę Affymetrix, Inc. i jest przez nią sprzedawane pod marką GeneChip. Przewiduje ono wykorzystanie oligonukleotydów, których synteza odbywa się już na macierzy. Płytką wykonaną jest ze szkła, ma mniejsze rozmiary i o wiele bardziej skomplikowaną strukturę. Do odczytu wartości wykorzystywany jest laser jednokanałowy. Generalnie eksperymenty są kosztowne, wymagają zakupu produktu spośród określonego w ofercie zakresu lub na drodze specjalnego zamówienia.

Ponieważ praca dotyczy głównie analizy otrzymanych danych, dlatego dalej pojęcie mikromacierzy będzie stosowane dla określenia ogólnej idei postępowania.

2.4. Schemat eksperymentu

Eksperyment wykonywany jest dla pewnej populacji komórek, których bieżący stan zamierzamy badać. Zasadniczo dla wykonania pomiaru ekspresji genów danej komórki wykorzystywana jest oddzielna mikromacierz. Przy zastosowaniu różnego znakowania

sekwencji możliwe jest dokonanie pomiaru dla większej liczby komórek przy wykorzystaniu jednej macierzy.



Rys. 2.4-1. Schemat pomiaru ekspresji genów komórki za pomocą mikromacierzy w technologii proponowanej przez Affymetrix [51]

Na rysunku 2.4-1 znajduje się schemat przykładowego sposobu pomiaru aktywności genów badanej komórki. Założono, że mikromacierz z sondami została przygotowana wcześniej, niezależnie od doświadczenia. Przedstawioną procedurę często utożsamia się z jej najistotniejszym etapem, jakim jest hybrydyzacja. W literaturze spotyka się także określenie próba lub przykład (ang. sample). Można powiedzieć, że eksperyment składa się z serii hybrydyzacji, po których następuje analiza danych pomiarowych i wizualizacja otrzymanych wyników.

Zanim przejdziemy do omówienia pojedynczej hybrydyzacji zaznaczmy, że w literaturze panuje pewna niejednoznaczność w określaniu sekwencji znajdujących się na płytce i sekwencji badanych. Przykładowo rozbieżne nazewnictwo używane jest w organizacjach KBRIN [53] i Affymetrix [51]. W pracy przyjmujemy nomenklaturę wykorzystywaną w Affymetrix, gdyż jest ona również zalecana przez Nature Genetics [56]. Zgodnie z nią sekwencja znana określana jest jako sonda (ang. probe), a sekwencja badana jako cel (ang. target). W rozpatrywanych przez nas technologiach sondy nanoszone są na płytkę, a cele pobrane z komórek umieszczane są w roztworze. Jest to istotne, gdyż w zastosowaniach spotyka się również odwrotny sposób rozmieszczenia obu zestawów łańcuchów, np. ONF (Oligonucleotide FingerPrinting).

Wybór odpowiedniej technologii dla eksperymentu znacząco wpływa na sposób pomiaru ekspresji. Pomimo to procedura ma zasadniczo podobny przebieg.

1. Przygotowanie sond

Na początku eksperymentu wybierany jest pewien zestaw genów danego organizmu. Zakres selekcji zależy od celu przeprowadzanego badania i może obejmować nawet cały genom.

Źródłem materiału genetycznego są dostępne biblioteki genowe. Dla każdej wybranej sekwencji syntetyzowane są sekwencje komplementarne, które nazywa się sondami. W zależności od specyfiki przyjętej technologii sondy mogą być sekwencjami cDNA lub oligonukleotydów.

2. Przygotowanie mikromacierzy

Etap ten może być przeprowadzony przez zespół prowadzący eksperyment lub wykonany przez specjalistyczną organizację zewnętrzną. Zbiory identycznych kopii sond są umieszczane w oddzielnych komórkach płytki. Komórki są adresowane poprzez numer wiersza i kolumny, co jest szczególnie istotne dla późniejszej analizy ekspresji konkretnych sekwencji. W zależności od typu macierzy sondy są przygotowywane niezależnie i następnie nakładane na płytkę lub są syntetyzowane in situ.

W większości przypadków z punktu widzenia analizy danych wynikowych najlepiej jest móc skojarzyć każdą komórkę mikromacierzy z dokładnie jednym genem. Niestety często uzyskanie takiej fragmentaryzacji materiału genetycznego nie jest możliwe i sekwencje składające się na pojedynczy gen znajdują się w różnych lokacjach. Ponadto ze względu na późniejszą weryfikację celowe jest umieszczanie materiału dotyczącego tego samego genu (replikantów) w kilku osobnych komórkach płytki. Sposób rozłożenia materiału genetycznego musi być odpowiednio uwzględniony po pomiarze intensywności ekspresji w celu wyznaczenia jej poziomów dla poszczególnych sekwencji.

3. Przygotowanie sekwencji badanych

W celu pomiaru intensywności ekspresji genów danej komórki w pierwszej kolejności należy wyprowadzić z jej środowiska całe transkrybowane matrycowe RNA. Następnie materiał jest odwrotnie transkrybowany do postaci cDNA. Zauważmy, że etap ten jest analogiczny do opisanego wcześniej klonowania materiału genetycznego, dokonywanego np. w celu budowy biblioteki cDNA.

Otrzymane sekwencje cDNA są specjalnie znakowane poprzez chemiczną modyfikację ich łańcuchów. Znakowanie nie ingeruje w pierwszorzędową strukturę kwasu. Proces może być realizowany w czasie odwrotnej transkrypcji lub już na otrzymanym cDNA. Pierwsze rozwiązanie jest zalecane ze względu na bardziej równomierną inkorporację barwnika przez cząsteczki.

Celem znakowania jest umożliwienie odnotowania przez urządzenia miernicze zajścia hybrydyzacji pomiędzy sekwencjami badanymi i sondami. Przykładowe znakowanie może być związane ze zjawiskami o charakterze fluorescencyjnym lub radioaktywnym.

Pojedyncza nić oznakowanego kwasu jest fragmentaryzowana na odcinki, które będą dołączać się do komplementarnych odcinków na płycie.

Ze względu na późniejszą weryfikację pomiaru często przygotowuje się również zestaw sekwencji kontrolnych, które są później wykorzystywane w procesie normalizacji.

4. Hybrydyzacja

Płytką z sondami jest umieszczana w środowisku sekwencji badanych. Może się to odbywać przez kontakt ze sporządzoną miksturą lub odseparowanymi grupami sekwencji. Hybrydyzacja jest przeprowadzana przy użyciu różnych technik, np. fluorocytoometrii lub elektroforezy. Na etapie tym dochodzi do powiązania się badanych sekwencji z komplementarnymi sondami. Połączenie dwóch odpowiadających sobie sekwencji genetycznych umożliwia późniejszą rejestrację użytego znacznika.

5. *Oczyszczenie płytki*

Mikromacierz po hybrydyzacji musi zostać przygotowana do pomiaru intensywności promieniowania. Przede wszystkim należy usunąć sekwencje badane, które nie uległy hybrydyzacji, gdyż zawarty w nich barwnik mógłby zaburzyć pomiar emisji pochodzącej z cząsteczek dwuniciowych. Płytką zostaje umieszczona w odpowiednim roztworze, następnie myta wielokrotnie i ostatecznie suszona.

6. *Odczyt poziomów ekspresji*

Duże stężenie wolnych sekwencji badanych w roztworze prowadzi do otrzymania w wyniku hybrydyzacji dużego stężenia połączonych helis w odpowiedniej komórce płytki. Fakt ten objawia się intensywniejszym oznakowaniem danej komórki.

W celu pomiaru intensywności promieniowania płytka jest skanowana przy pomocy odpowiedniego lasera. Odczytane wartości są zapisywane zgodnie z przyjętą jednostką i czułością urządzeń pomiarowych, a następnie przyporządkowywane opisom sond, które znajdowały się w określonych komórkach.

Odczyt wyników ze wszystkich płytek serii kończy część związaną z pomiarem poziomów ekspresji. Kolejne etapy dotyczą przetwarzania danych wynikowych i są silnie związane z celem całego doświadczenia. Dla poprawnej interpretacji pomiarów istotne jest prowadzenie szczegółowego protokołu laboratoryjnego oraz zapis wyników pośrednich, np. zeskanowane obrazy płytek. W ten sposób możliwe jest powtórzenie analizy poczynając od wybranego etapu.

Większość ścieżek analizy danych rozpoczyna się czyszczeniem i normalizacją danych. Czyszczenie odpowiada za zgrubną kontrolę poprawności danych pomiarowych. Zawiera między innymi eliminację błędów grubych oraz oznakowanie wartości brakujących. Normalizacja wykorzystuje informacje kontrolne do odpowiedniej modyfikacji rozkładów wartości pochodzących z każdej płytki. Umożliwia ona porównywanie wyników otrzymanych w kolejnych hybrydyzacjach. Proces ten jest szczegółowo omówiony w rozdziale 2.6.

Po wstępnej obróbce danych następuje bardziej zaawansowana analiza właściwa, która często kończy się graficzną prezentacją wyników.

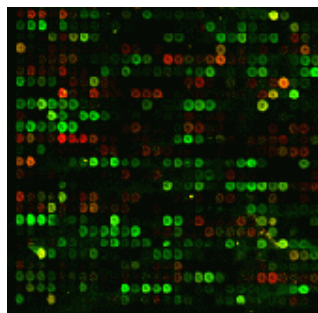
2.5. Pomiar poziomu ekspresji

Eksperymenty wykonywane są w różnych warunkach, przy użyciu różnego rodzaju mikromacierzy oraz sprzętu pomiarowego. Z tego względu niezbędne jest podejście, które pozwoli na jakiegokolwiek porównanie otrzymywanych wyników.

Ogólnie zastosowanie mikromacierzy pozwala na pomiar wartości poziomu ekspresji genów. Ze względu na brak standardu dla sposobu prowadzenia eksperymentów najczęściej przedmiotem analizy nie są otrzymywane wartości bezwzględne, lecz wartości odniesione do określonego poziomu.

W praktyce postulat ten realizowany jest poprzez jednoczesną hybrydyzację na pojedynczej macierzy dwóch różnie oznakowanych zestawów próbek. Oba zestawy dotyczą tego samego zakresu genomu. Jeden ma charakter kontrolny i jest tak przygotowany, aby dawać w przybliżeniu stały poziom ekspresji dla wszystkich genów. Drugi jest właściwym zestawem badanych. Ze względu na różne znakowanie aparatura pomiarowa może zmierzyć ekspresję osobno dla każdego zestawu. Na podstawie tych wartości dla każdego genu wyznaczany jest względny poziom ekspresji sekwencji badanych w stosunku do ekspresji sekwencji kontrolnych. Zauważmy, że jednoczesność pomiaru dla obu zestawów jest szczególnie istotna

dla minimalizacji błędu związanego z różnymi otoczeniami pomiarowymi. Z tego względu zestaw kontrolny stosuje się w każdej hybrydyzacji eksperymentu.



Rys. 2.5-1. Fragment zeskanowanej mikromacierzy zawierającej pełny genom drożdży

2.6. Normalizacja

Pomiar intensywności promieniowania emitowanego przez wykorzystany w eksperymencie materiał znakujący obarczony jest pewnym błędem. W przypadku pomiaru ekspresji przy użyciu mikromacierzy błąd ten może osiągać wysoki poziom. Składa się na niego wiele czynników, wśród których należy wymienić charakterystykę zastosowanej płytki, cechy materiału znakującego (wrażliwość na ciepło i światło, okres połowicznego rozpadu), równomierność namnażania i oznakowania sekwencji, przypadkowość procesu hybrydyzacji oraz parametry skanera mierzącego promieniowanie.

Rozpatrzmy eksperyment wykorzystujący dwa znakowania, jedno dla sekwencji badanych (R) i jedno dla kontrolnych (G). Celem etapu normalizacji jest minimalizacja błędu systematycznego towarzyszącego pomiarowi względnego poziomu ekspresji genów. Dalej będziemy zakładać, że dla każdego genu posiadamy parę pomiarów emisji pochodzących z obu znakowań oraz informację, w którym śladzie głowicy nakładającej materiał genetyczny znajduje się komórka płytki zawierająca gen. Prezentowane transformacje dotyczyć będą zbioru wartości $\log(R/G)$ dla wszystkich genów mikromacierzy.

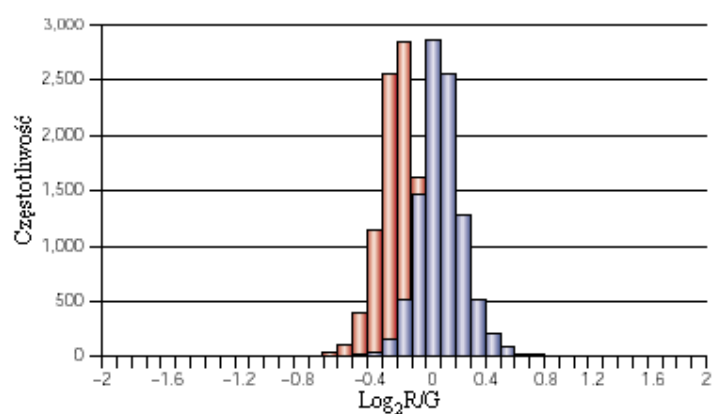
Zamierzona normalizacja jest silnie powiązana ze sposobem prowadzenia eksperymentu. Zgodnie z [50] metoda normalizacji może być oparta o wszystkie geny znajdujące się w macierzy, o geny ulegające stałej ekspresji (ang. housekeeping genes) lub o specjalnie umiejscowione sekwencje kontrolne.

Proces normalizacji może dotyczyć jednej lub wielu macierzy. Wśród metod stosowanych dla pojedynczej macierzy najpopularniejsza jest normalizacja globalna. Opiera się ona na założeniu, że stosunek intensywności obu barwników dla każdego genu jest stały. Środek rozkładu wartości jest przesuwany na 0. Za środek rozkładu przyjmuje się najczęściej średnią lub medianę. Bardziej zaawansowane metody normalizujące położenie rozkładu względnych poziomów ekspresji wykorzystują regresję liniową. W tym przypadku istotne jest dobranie odpowiedniego modelu statystycznego. Propozycja w zakresie tego podejścia zawarta jest w pracy [26].

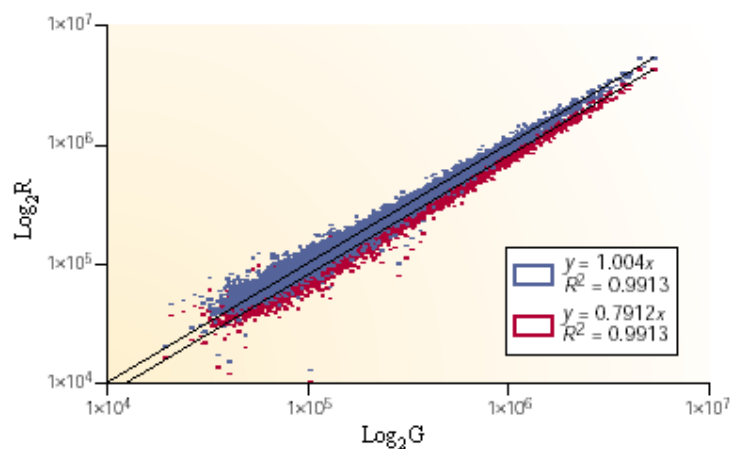
Przykład ([40])

Rozważamy eksperyment, w którym dla kanału R obserwujemy generalnie niższą intensywność niż dla kanału G. Na rysunku 2.6-1 znajduje się histogram przedstawiający częstotliwość występowania w danych wynikowych odpowiednich wartości $\log(R/G)$. W związku z nierówną intensywnością kanałów środek histogramu jest przesunięty w lewo w stosunku do punktu 0. Zastosowana została normalizacja położenia, która przesunęła rozkład wartości tak, aby jego środek pokrył się z punktem zero. Źródło nie podaje, czy za środek rozkładu została przyjęta średnia, czy mediana. Obserwowalna zamiana kształtu rozkładu jest artefaktem procedury tworzenia histogramu wymagającej ustalenia przyporządkowania wartości do określonych przedziałów.

Przeprowadzona transformacja została dodatkowo zilustrowana na rysunku 2.6-2 zawierającym wykres zależności pomiędzy logarytmami intensywności emisji w obu kanałach.



Rys. 2.6-1. Histogram dla rozkładu logarytmu stosunku intensywności kanału R do G. Kolor czerwony odpowiada rozkładowi pierwotnemu, a niebieski – znormalizowanemu



Rys. 2.6-2. Wykres punktowy zależności logarytmów intensywności w kanałach R i G. Kolor czerwony odpowiada rozkładowi pierwotnemu, a niebieski – znormalizowanemu

Specyfika technologii mikromacierzy wymaga wykształcenia metod uwzględniających najważniejsze źródła błędów. Jednym z nich jest wzięcie pod uwagę sposobu nakładania materiału genetycznego na płytkę. Uzasadnione jest oddzielne traktowanie prostokątnych obszarów płytki, które odpowiadają śladom głowicy drukującej. Dla każdej z tych grup może być przeprowadzona oddzielna normalizacja położenia rozkładu wartości odpowiadających zawartych w niej genom. Aby otrzymać wartości znormalizowane dla całej macierzy należy przeprowadzić normalizację skali pomiędzy obszarami. Najczęściej opiera się ją o założenie, że wartości w każdym obszarze mają rozkład normalny o wyliczanych dla danego obszaru parametrach.

Normalizacja wielu macierzy jest wymagana, gdy chcemy mieć możliwość porównywania wyników otrzymywanych przy kolejnych hybrydyzacjach. Postępowanie jest analogiczne do normalizacji uwzględniającej obszary macierzy. W pierwszym kroku należy przeprowadzić osobną normalizację każdej z rozpatrywanych macierzy. Następnie przeprowadzana jest normalizacja skali przy założeniu, że wartości z każdej macierzy mają odpowiedni rozkład normalny.

2.7. Rodzaje analizy danych

Właściwa analiza danych może być związana zarówno z uczeniem się bez nadzoru, jak i z nadzorem. W kontekście badania ekspresji genów rozróżnienie to związane jest z pytaniem, czy przygotowując doświadczenie dysponujemy wiedzą na temat klas badanych organizmów. Klasa jest tu ściśle związana z tematyką eksperymentu, może ona przykładowo dotyczyć typów leczenia, jakie przeszły poszczególne organizmy, posiadanych przez nie schorzeń itp. Celem jest zbadanie powiązania informacji o ekspresji genów z konkretnymi zjawiskami fenotypowymi. Zgodnie z założeniami z rozdziału 1.2 w pracy skupimy się wyłącznie na przypadku, gdy klasa organizmów jest znana, czyli na uczeniu z nadzorem.

Zgodnie z powyższym możemy wyróżnić dwa główne typy eksperymentów [31]. Pierwszy typ związany jest z jednoczesnym pomiarem genów dla serii różnych warunków, np. dla kolejnych stadiów cyklu życiowego komórki lub stanu po określonej stymulacji. W ten sposób można śledzić trendy i regularności w zachowaniu poszczególnych genów. Rezultaty mają charakter temporalny.

Drugi typ obserwuje geny różnych komórek, które znajdują się w tym samym środowisku. Badania mogą dotyczyć populacji komórek organizmu danego gatunku, której osobniki przeszły odmienne procesy leczenia, zostały skażone toksynami lub cierpią na pewne schorzenia, np. nowotwory. Uzyskane informacje dotyczą relacji pomiędzy genami a określonym zestawem cech fenotypowych, które mogą być wykorzystane do klasyfikacji nowych komórek oraz do identyfikacji genów o dużym znaczeniu diagnostycznym. Dane wynikowe mają charakter przestrzenny.

2.8. Dane dostępne publicznie

Rezultaty eksperymentów dostępne są na stronach związanych z placówkami badawczymi i gromadzone w bazach mikromacierzy.

Doświadczenia w wielu placówkach prowadzone są zgodnie z przyjętymi na potrzeby projektu założeniami. Nie istnieją ogólnie przyjęte zasady prowadzenia eksperymentu, dokumentowania wyników oraz ich wstępnego przetworzenia przed bardziej zaawansowaną analizą. Pomimo wielu podobieństw w schematach eksperymentów i stosowania podobnych

algorytmów czyszczenia i normalizacji publiczne dostępne dane mogą różnić się w sposób znaczący. Najczęściej prezentowane są wyniki eksperymentów, które nie przeszły pełnego przetwarzania wstępnego.

W publikowanych rezultatach napotykamy na następujące problemy:

- Wyniki podane dla sekwencji, które znajdowały się na macierzy zamiast dla faktycznych genów badanego organizmu
- Brak wartości poziomów ekspresji genów dla części przykładów
- Nie uwzględnienie replikantów
- Nieodniesienie wartości badanych do kontrolnych
- Nieznormalizowanie wartości
- Niekompletna dokumentacja eksperymentu i zamieszczanych danych

Przed rozpoczęciem właściwej analizy niezbędne jest doprowadzenie danych do odpowiedniej postaci. Wymaga to dokończenia etapu przetwarzania wstępnego.

2.9. Przykład eksperymentu

W rozdziale tym przyjrzymy się eksperymentowi przeprowadzonego przez zespół badawczy z University of Toronto Health Network. Dotyczy on zbadania możliwości klasyfikacji tkanek zdrowych i zrakowaciałych za pomocą profili ekspresji genów. Doświadczenie zostało dokładnie omówione w pracy [46].

Komórki pobrano od 39 pacjentów chorych na raka, którzy przeszli leczenie i znajdowali się pod obserwacją. Po roku u 24 osób stwierdzono nawrót nowotworu, podczas gdy 15 okazało się zdrowych.

W doświadczeniu wykorzystano mikromacierze cDNA z ludzkim genomem (ok. 18100 sekwencji). Pojedyncza próba dotyczyła jednego pacjenta. Z komórek pacjenta został pobrany matrycowy RNA, z którego otrzymano cDNA za pomocą odwrotnej transkrypcji. Poza sekwencjami badanymi wykorzystane zostały sekwencje kontrolne o możliwie stałej ekspresji dla całego genomu. Sekwencje badane cDNA zostały oznakowane barwnikiem Cy5 (czerwony, R), a kontrolne barwnikiem Cy3 (zielony, G). Cele zostały poddane hybrydyzacji na mikromacierzach. Emisja związana ze znakowaniem została zmierzona za pomocą skanera Genepix 4000. Otrzymane wartości bezwzględne znormalizowano. W macierzy wynikowej dla każdego genu podano wartość $\log_2 R/G$.

Dane bezpośrednio otrzymane z pomiaru charakteryzowały się dużym odsetkiem wartości brakujących. Do dalszej analizy został wybrany podzbiór 2880 genów (wg artykułu 2899), dla których odczytano poziom ekspresji dla co najmniej 80% przypadków oraz dla których przynajmniej w dwóch przypadkach stwierdzono co najmniej 4-krotną zmianę wartości badanej do kontrolnej.

Zaawansowana analiza danych związana była z uczeniem bez nadzoru. Profile ekspresji pacjentów zostały poddane grupowaniu hierarchicznemu. Otrzymane dwa główne klastry odpowiadały dwóm klasom (zdrowy, chory), do których przydzielono pacjentów po rocznej obserwacji. Fakt ten uzasadnia podejście molekularne do klasyfikacji tkanek z nowotworami. Główne klastry wskazały również geny, które mogą być związane ze schorzeniami onkologicznymi. Geny te mogą być dobrymi wskaźnikami do klasyfikacji tkanek.

Dane wynikowe są dostępne publicznie na stronie [55]. Autorzy udostępniają zarówno dane z pełnym zestawem genów, jak i po wstępnej eliminacji genów z silnie niekompletnymi

poziomami ekspresji dla wszystkich przykładów. Wszystkie pliki zostały zapisane w formie tekstowej i w formacie programu Microsoft Excel (XLS).

Wyniki są zorganizowane następująco:

- Plik zawierający tabelę ekspresji, której wiersze opisane są symbolami sekwencji genetycznych, a kolumny identyfikatorami przykładów (tabela 2.10-1)
- Plik zawierający tabelę ekspresji po eliminacji genów
- Plikiem zawierający powiązanie każdego identyfikatora przykładu z rozszerzonymi danymi na temat pacjenta. Do danych tych należą szczegóły prowadzonej obserwacji oraz ostatecznie wybrana klasa decyzyjna (tabela 2.10-2)

UNIQID	NAME	B_020T_218091_ADC_2B rec	B_022T_218144_LCUC_1B rec
588029	588029:Hs.79:ACY1	0,195	0,375
190929	190929:Hs.247565:RHO	0,15	0,21
246550	246550:Hs.293548:not available		

Tab. 2.10-1. Fragment arkusza opisującego ekspresję genów w badanych przypadkach

TB-id	Histopathology	Recur	~% T c	Stage	Surg date	LFU	Status	Rec Date
P038T1	ADC	N	70%	1A	06-sty-97	02-kwi-97	A	-
P153T1	ADC	N	80%	1A	21-maj-98	07-cze-99	A	-
P188T2	ADC	N	50%	1A	26-lis-98	15-maj-01	A	-
P174T1	carcinoid	N	50%	1A	22-wrz-98	19-kwi-00	A	-
P149T1	SQCC	N	80%	1A	23-kwi-98	01-maj-00	A	-
P204T2	SQCC	N	40%	1A	15-lut-99	18-sty-01	A	-
B024T1	ADC	Y	60%	1B	05-kwi-95	01-mar-00	A	03-maj-96
B051T1	ADC	Y	50%	1B	01-sie-95	14-kwi-97	D	24-mar-97

Tab. 2.10-2. Fragment arkusza opisującego badane przypadki

Niestety nawet dane po przetworzeniu wciąż zawierają pewną liczbę wartości brakujących. W przypadku wielu bardziej zaawansowanych metod, a także w przypadku prezentowanej w pracy ścieżki analizy niezbędne jest uzupełnienie niekompletnych danych.

2.10. Standaryzacja opisu

Rozproszony charakter badań towarzyszący szybkiemu rozwojowi kolejnych gałęzi bioinformatyki rodzi problemy standaryzacyjne. Stanowią one punkt wyjścia do uporządkowania skojarzonych z eksperymentami danych.

W przypadku najstarszej gałęzi, jaką są badania sekwencji genetycznych dość wcześnie zaczęły powstawać bazy danych skupione wokół placówek naukowych. Najważniejszymi wśród nich są EMBL w Europie i GenBank w Stanach Zjednoczonych. Obecnie proces standaryzacji danych dotyczących sekwencji genów jest mocno zaawansowany. Dokumentacja wyników otrzymywanych w kilku dominujących organizacjach ujęta jest w bazy, które intensywnie ze sobą współpracują. W tej dziedzinie współistnieje, co prawda kilka formatów związanych z najważniejszymi organizacjami i narzędziami (GenBank, EMBL, Clustal i in.), jednak są one wzajemnie respektowane i podlegają ciągłej

konwergencji. Akceptację zdobył również tekstowy standard BSML (Bioinformatic Markup Language), syntaktycznie oparty o XML. Pierwotnie był on związany właśnie z opisem sekwencji genetycznych i pomimo ogólnego charakteru, jaki zawarty jest w jego nazwie pozostaje wciąż związany z tą tematyką. Bogate informacje na temat standardu można pobrać ze strony [52]. Warto dodać, że o dynamicznym tempie rozwoju badań nad sekwencjami zdecydował w większości publiczny charakter zasobów oraz podstawowych narzędzi do ich obsługi.

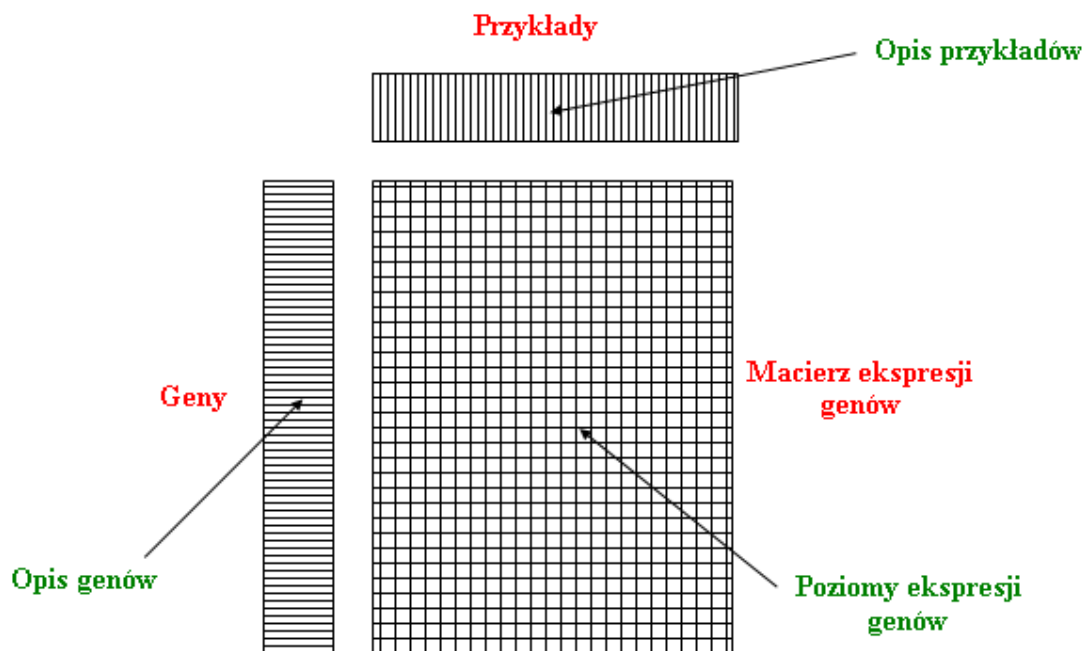
Badania nad ekspresją genów są o wiele młodsze, co pociąga za sobą brak jednolitości dokumentowania danych. Obecnie dostępne opisy mają postać dokumentów tekstowych opracowanych zgodnie z własnymi założeniami grup badawczych. Powstają także pierwsze bazy danych związane silnie z aktywnymi w dziedzinie placówkami badawczymi.

Podstawowym celem jest doprowadzenie do powstania powszechnego repozytorium wyników eksperymentów. W tej kwestii na pierwszym miejscu stoi problem ujednolicenia wymagań dotyczących opisu i stworzenia według nich struktury bazy danych.

2.11. Koncepcja bazy danych mikromacierzy

Istnieją dwie zasadnicze grupy repozytoriów danych na temat ekspresji genów. Pierwsza grupa zawiera bazy określane tradycyjnie mianem baz ekspresji genów (ang. gene expression database), które przechowują wyniki otrzymywane za pomocą starszych technologii, takich jak SAGE. Druga związana jest wyłącznie z eksperymentami wykorzystującymi mikromacierze i obejmuje bazy danych nazywane bazami mikromacierzy (ang. microarray database). Duża różnica pomiędzy najnowszą technologią a rozwiązaniami wcześniejszymi wpływa na odmienną strukturę baz obu grup. W rozdziale tym rozważymy ogólne podejście do tworzenia bazy danych na temat ekspresji genów otrzymywanych z doświadczeń w technologii mikromacierzy.

Dane z pojedynczej hybrydyzacji można interpretować jako wektor wartości uzyskanych dla określonych genów. Same hybrydyzacje mogą pochodzić z wielu różnych eksperymentów. Sytuacja taka sugeruje dwuwymiarową strukturę bazy danych. Baza zbudowana jest wokół zbioru wartości poziomów ekspresji, a jej wymiarami są geny i hybrydyzacje. Ponieważ eksperymenty prowadzone są dla różnych gatunków organizmów o określonych genomach, więc macierz ekspresji genów jest rzadka. Ideowy schemat struktury został przedstawiony na rysunku 2.11-1.



Rys. 2.11-1. Konceptualna struktura bazy danych na temat ekspresji genów

Zastanówmy się, jakie dane powinny być przechowywane dla każdego wymiaru bazy, aby mogła ona udzielać użytecznych informacji na temat zachowania genów w określonych warunkach.

Dobry opis genów zapewniają istniejące zewnętrzne bazy sekwencji, więc w tym zakresie możemy ograniczyć się do wskazania unikalnych identyfikatorów. W przypadku prowadzonych eksperymentów istotny jest jednoznaczny opis kontekstu laboratoryjnego dla każdej hybrydyzacji. Niestety nie dysponujemy gotowymi bazami wzorców prowadzenia doświadczeń i możemy powoływać się jedynie na źródła poruszające analizowane gatunki organizmów oraz dokumentację używanego sprzętu i oprogramowania. Z wyjątkiem publikacji taksonomicznych dane te nie znajdują się w publicznych repozytoriach i należałoby je dołączać do opisu każdego przykładu.

Drugi problem stanowi unifikacja jednostek, w których mierzy się poziomy ekspresji genów. Wynika on głównie z różnych sposobów znakowania sekwencji i pomiaru ich poziomu ekspresji. Dodatkowo na rynku brak wspólnych standardów określających parametry mikromacierzy. Obecnie konieczne jest dołączanie do eksperymentu serii zeskanowanych obrazów pozwalających na przyjęcie różnych sposobów pomiaru intensywności. Pomijając względy praktyczne otwarta pozostaje kwestia, czy w ogóle istnieje wspólna miara, która pozwoliłaby jednoznacznie reprezentować poziom ekspresji genów i czy technologia mikromacierzy pozwala na pomiar wartości tej miary. Podsumowując, informacja na temat sposobu pomiaru poziomu ekspresji musi zostać dołączona do opisu wymiaru przykładów.

Dalszy rozwój technologii mikromacierzy pozwoli na określenie dokładnych parametrów błędu pomiaru dla każdej komórki. Informacja ta może zostać dołączona do każdego przechowywanego poziomu ekspresji.

Listę wielu dostępnych baz mikromacierzy można znaleźć na stronie [59].

2.12. Specyfikacja MIAME

Proces standaryzacyjny w dziedzinie ekspresji genów związany jest z międzynarodowym stowarzyszeniem MGED Society (Microarray Gene Expression Data Society). W listopadzie 1999 z jego inicjatywy powstała grupa odpowiedzialna za opracowanie specyfikacji pod nazwą MIAME (Minimum Information About a Microarray Experiment). Obecnie jest dostępna wersja 1.1 Draft 6.

Opracowywana specyfikacja ma za zadanie umożliwić budowę powszechnego formatu wymiany oraz utworzenie rozproszonego systemu baz danych. W bieżącym stanie prac ma ona charakter nieformalny. Jest zbiorem podstawowych wymogów oraz zaleceń dotyczących sposobu dokumentowania eksperymentów.

Z punktu widzenia syntaktyki dopuszczany jest zapis wolnym tekstem przy zachowaniu pewnych wytycznych. Opis parametrów powinien być w notacji trypletowej o postaci:

(nazwa parametru, wartość, źródło)

Przykład

Poniższy tryplet może opisywać pewną komórkę organizmu:
(typ komórki, komórka mięśniowa, Anatomia człowieka (Jose Parramon, „Galaktyka”))

Najistotniejszą częścią specyfikacji jest ogólny plan dokumentowania eksperymentu z wyszczególnieniem informacji, które powinien zawierać.

Schemat obejmuje kolejno:

- Opis eksperymentu
 - Ogólne dane (typ, testowane parametry, lista organizmów)
 - Lista przykładów, tablic, hybrydyzacji
 - Wskaźniki jakości
 - Dokładny opis
- Opis macierzy
 - Dokładny opis każdej mikromacierzy i jej komórek
 - Przyporządkowanie identyfikatorów mikromacierzom
- Opis przykładów
 - Użyty organizm (zgodnie z taksonomią NCBI)
 - Informacje o stanie organizmów (faza rozwoju, leczenie itp.)
 - Protokół laboratoryjny z przygotowania roztworu do hybrydyzacji
 - Protokół laboratoryjny z znakowania sekwencji
- Opis hybrydyzacji
 - Protokół laboratoryjny (procedura i parametry)
- Wyniki i opis pomiarów
 - Zeskanowane obrazy wyników hybrydyzacji (format TIFF, DAT)
 - Dane przejściowe po analizie obrazów
 - Dane końcowe po uwzględnieniu replikantów
- Opis normalizacji
 - Podejście i użyty algorytm
 - Wyniki normalizacji

Bieżących informacji na temat MIAME dostarcza strona [54].

3. Formalny opis danych

3.1. Wprowadzenie

Badając fragment rzeczywistości musimy zdecydować się na przyjęcie pewnego ścisłego modelu dla umożliwienia abstrakcyjnej reprezentacji realnych pojęć. Model ten pozwala na formułowanie i weryfikację hipotez, które później mogą być aplikowane z powrotem do rzeczywistości.

W pierwszej kolejności należy zdecydować się na wybór pewnej przestrzeni obiektów odpowiadających realnym faktom. Wybór ten powinien uwzględniać wszystkie możliwe fakty. Po drugie trzeba wybrać przestrzeń atrybutów, które będą określały cechy obiektów.

W rozpatrywanym zagadnieniu wiedza dotyczy organizmów żywych i jest oparta o analizę materiału genetycznego oraz obserwację cech niegenetycznych, np. fenotypowych. Obiekty reprezentują konkretne komórki, tkanki lub proste organizmy, które są rozpatrywane w ścisłym związku ze swoim bieżącym stanem. Dla przejrzystości bieżącej pracy przyjmujemy, że obiekt reprezentuje organizm.

W pierwszej kolejności obiekt określony jest przez klasę organizmu w skali makro, która może wynikać z etapu cyklu życiowego, przeprowadzonej wcześniej terapii lub charakteru otaczającego środowiska. Atrybut ten ma charakter nominalny. Przykładowo może określać podział na komórki zdrowe i zrakowaciałe albo organizmy skażone toksynami A, B lub C. Należy zaznaczyć, że pomimo kluczowego znaczenia atrybutu decyzyjnego dla dalszej analizy zmuszeni jesteśmy polegać na jego zgrubnym określeniu. Jego wybór bezpośrednio uzasadniony jest celem prowadzonej analizy, a ustalenie przeciwdziałziny wynika z bieżącej wiedzy na temat zagadnienia. Może to doprowadzić do wspólnej identyfikacji znacząco różnych przypadków. Przykładowo dwa zrakowaciałe organizmy mogą dodatkowo różnić się pod względem charakteru nowotworu, co skutecznie zaburza interpretację ich stanów genetycznych. Możliwe, że przy niewielkiej liczbie obiektów związanych z eksperymentem nie będzie można znaleźć klarownej hipotezy.

Ponadto organizm charakteryzowany jest poprzez szereg atrybutów odpowiadających aktywności jego poszczególnych genów. Dany model precyzuje rozpatrywany zakres genów. Uściślając, dla każdego z tych genów istnieje atrybut określający poziom ekspresji tego genu.

Eksperyment obejmuje zbadanie pewnego zbioru możliwych instancji organizmów poprzez określenie ich stanu. W ramach modelu jest on opisywany przez zbiór obiektów o wartościach atrybutów odpowiadających pomiarom parametrów składających się na stan. Reprezentacja ta odpowiada strukturze tablicy decyzyjnej, której atrybutami warunkowymi są poziomy ekspresji, a atrybutem decyzyjnym – klasa komórki.

Trzeba mieć świadomość, że jesteśmy w stanie obserwować jedynie podzbiór możliwych organizmów. Ze względu na słuszność wysnuwanych wniosków ważne jest, aby był on reprezentatywny.

3.2. Definicje

W rozdziale tym podajemy definicje podstawowych pojęć wykorzystywanych w kolejnych rozdziałach. Tutaj także wprowadzony jest przewodni przykład, za pomocą którego zilustrowana zostanie proponowana w pracy ścieżka analizy danych.

Przykład

Zajmijmy się analizą danych pochodzących z hipotetycznego eksperymentu dotyczącego badania nowotworów. Rozważmy prosty organizm, którego genom złożony jest z 8 genów identyfikowanych przez symbole $G_1, \dots, G_8$. Za pomocą technologii mikromacierzy badany jest poziom ekspresji wszystkich genów dla komórek pochodzących z tkanek tego samego typu z kilku wybranych organizmów. Eksperyment obejmuje serię 8 hybrydyzacji, oznaczonych przez $H_1, \dots, H_8$, przy czym 4 pierwsze hybrydyzacje przeprowadzane są na materiale pobranym z komórek zdrowych, a 4 pozostałe na materiale z komórek zrakowaciałych. Fakt ten odnotowany jest dla każdej hybrydyzacji poprzez klasę Z(Zdrowy) lub R(Rak).

Założmy dalej, że aparatura pomiarowa podaje wyniki dla każdej komórki macierzy w postaci znormalizowanej, przy czym skala jest przedziałem $(0,10)$.

Wyniki dotyczące poziomów ekspresji genów dla kolejnych hybrydyzacji zgromadzone zostały w tabeli 3.2-1. Sposób przedstawienia danych jest powszechnie przyjęty przy zapisie wyników eksperymentów z użyciem mikromacierzy.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
G_1	1,01	1,74	1,52	1,02	3,17	4,12	8,17	9,65
G_2	0,21	0,11	0,12	0,03	1,00	1,53	0,99	0,89
G_3	2,41	3,12	5,22	6,33	0,37	0,36	0,91	0,51
G_4	5,51	8,34	3,32	6,02	1,17	3,12	5,17	1,65
G_5	6,44	6,26	4,12	8,02	3,02	3,18	2,11	3,33
G_6	9,14	6,56	1,38	1,66	8,67	4,52	9,21	9,25
G_7	4,41	5,17	7,16	8,36	4,22	7,16	8,90	5,16
G_8	2,41	3,12	5,22	6,33	2,56	5,22	6,98	4,21
d	Z	Z	Z	Z	R	R	R	R

Tab. 3.2-1. Tablica wyników dla eksperymentu zawierająca poziomy ekspresji genów

Dla przejrzystości wprowadzone symboliczne oznaczenia pojęć rzeczywistych zostaną bez żadnych zmian zastosowane dla pojęć abstrakcyjnych opisu formalnego.

Definicja 3.2-1

Uniwersum (U) oznacza zbiór obiektów.

Przykład

$$U = \{H_1, \dots, H_8\}$$

Definicja 3.2-2

Atrybut oznacza funkcję $a : U \rightarrow V_a$.

Przykład

Dla każdego $i=1, \dots, 8$ mamy atrybuty określające poziom ekspresji i -tego genu $G_i : U \rightarrow V_i$.

Przeciwdziedzina każdego z tych atrybutów jest ustalona jako przedział, którego lewym krańcem jest najmniejsza wartość ekspresji odpowiadającego mu genu, a prawym krańcem – największa.

Atrybut	Przeciwdziedzina atrybutu V_i
G_1	$\langle 1,01;9,65 \rangle$
G_2	$\langle 0,03;1,53 \rangle$
G_3	$\langle 0,36;6,33 \rangle$
G_4	$\langle 1,17;8,34 \rangle$
G_5	$\langle 2,11;8,02 \rangle$
G_6	$\langle 1,38;9,25 \rangle$
G_7	$\langle 4,22;8,9 \rangle$
G_8	$\langle 2,41;6,98 \rangle$

Tab. 3.2-2. Przeciwdziedziny atrybutów opisujących ekspresję genów

Atrybut określający klasę jest równy $d : U \rightarrow \{Z, R\}$.

Definicja 3.2-3

Atrybut a jest:

- **dyskretny**, jeśli V_a jest dyskretny.
- **nominalny**, jeśli V_a jest nieuporządkowanym zbiorem wartości dyskretnych
- **porządkowy**, jeśli V_a jest uporządkowanym zbiorem wartości dyskretnych
- **numeryczny**, jeśli $V_a \subset \mathbb{R}$
- **ciągły**, jeśli $V_a \subset \mathbb{R}$ i V_a nie jest dyskretny
- **skończony**, jeśli V_a jest skończony.

Przykład

Atrybuty określające geny są numeryczne, ciągłe i skończone.

Atrybut określający klasę jest dyskretny, nominalny i skończony.

Definicja 3.2-4

System informacyjny (IS) oznacza parę (U, A) , taką że:

- U jest niepustym, skończonym zbiorem obiektów
- A niepustym, skończonym zbiorem atrybutów

Przykład

Zgodnie z poprzednimi rozważaniami system informacyjny zawierający informacje pochodzące z eksperymentu ma postać: $(\{H_1, \dots, H_{10}\}, \{G_1, \dots, G_8, d\})$.

Definicja 3.2-5

Tablica decyzyjna (DT) oznacza czwórkę (U, A, C, D) , taką że:

- (U, A) jest systemem informacyjnym
- $\{C, D\}$ jest podziałem A

Definicja 3.2-6

Niech dana będzie tablica decyzyjna (U, A, C, D) .

Atrybut a jest:

- **warunkowy**, jeśli $a \in C$
- **decyzyjny**, jeśli $a \in D$

Najczęściej rozpatruje się tablice decyzyjne, dla których $|D|=1$. Taki charakter mają na przykład tablice wykorzystywane do problemów klasyfikacji. Z tego względu będziemy dalej rozpatrywać wyłącznie tablice decyzyjne postaci $(U, A, C, \{d\})$, gdzie d jest jedynym atrybutem decyzyjnym.

Dla dalszej wygody wprowadzam dla tablicy decyzyjnej $(U, A, C, \{d\})$ następujące oznaczenia:

- $m=|U|$
- $n=|C|$
- $p=|Vd|$
- $U=\{u_1, \dots, u_m\}$
- $C=\{a_1, \dots, a_n\}$.

Przykład

Analogicznie do zbudowanego systemu informacyjnego w celu zapisania wyniku eksperymentu możemy wykorzystać tablicę decyzyjną. Charakter atrybutów podsuwa podział zbioru atrybutów na atrybuty warunkowe $\{G_1, \dots, G_8\}$ opisujące pomiary poziomu ekspresji poszczególnych genów oraz na jednoelementowy zbiór atrybutów decyzyjnych $\{d\}$ zawierających atrybut określający klasę komórki.

Tablica decyzyjna przyjmuje postać: $DT_1 = (\{H_1, \dots, H_{10}\}, \{G_1, \dots, G_8, d\}, \{G_1, \dots, G_8\}, \{d\})$. Dla przejrzystości atrybuty systemu informacyjnego lub tablicy decyzyjnej możemy przedstawić w postaci tabeli.

	G ₁	G ₂	G ₃	G ₄	G ₅	G ₆	G ₇	G ₈	d
H ₁	1,01	0,21	2,41	5,51	6,44	9,14	4,41	2,41	Z
H ₂	1,74	0,11	3,12	8,34	6,26	6,56	5,17	3,12	Z
H ₃	1,52	0,12	5,22	3,32	4,12	1,38	7,16	5,22	Z
H ₄	1,02	0,03	6,33	6,02	8,02	1,66	8,36	6,33	Z
H ₅	3,17	1,00	0,37	1,17	3,02	8,67	4,22	2,56	R
H ₆	4,12	1,53	0,36	3,12	3,18	4,52	7,16	5,22	R
H ₇	8,17	0,99	0,91	5,17	2,11	9,21	8,90	6,98	R
H ₈	9,65	0,89	0,51	1,65	3,33	9,25	5,16	4,21	R

Tab. 3.2-3. Tablica decyzyjna DT_1 opisująca dane wynikowe eksperymentu

W tabeli 3.2-3 wyróżniono zbiory atrybutów warunkowych i decyzyjnych. Tablica została zapisana w tradycyjnej postaci, dla której wiersze odpowiadają obiektom, a kolumny atrybutom. W stosunku do tablicy z tabeli 3.2-1 jest ona transponowana. W dalszej części pracy pierwotne dane wynikowe eksperymentu zapisane będą za pomocą tablicy decyzyjnej. Zgodnie z oznaczeniami zmienne przyjmują następujące wartości: $m=8$, $n=8$, $p=2$.

4. Wyznaczanie zbioru genów znaczących

4.1. Wprowadzenie

Definiując model decydujemy się na wybór przestrzeni atrybutów, które mogą wносить istotne dla zagadnienia informacje. Wiele względów przemawia za zawężeniem tej przestrzeni do zbioru atrybutów faktycznie związanych lub związanych odpowiednio silnie z badanymi hipotezami. Będzie on dalej nazywany zbiorem znaczącym.

Dla poprawności całej analizy kluczowy jest wybór rozpatrywanej przestrzeni, jako górnego ograniczenia na zbiór znaczący. Pomijając część czynników może się okazać, że zaproponowany model nie przystaje do rzeczywistości. Niestety nie posiadamy apriorycznej wiedzy na temat ogółu możliwych czynników i dlatego musimy zdecydować się na pewne przybliżenia.

Pierwsze trywialne oszacowanie odpowiada wyborowi wszystkich czynników mierzalnych ilościowo lub jakościowo. Może ono zostać skutecznie poprawione przy uwzględnieniu zdobytej wcześniej wiedzy na temat badanego zagadnienia.

Szerszego komentarza wymaga dobór atrybutów dla analizy danych o ekspresji. Eksperyment jest dalej reprezentowany przez tablicę decyzyjną. W tym wypadku mamy do ustalenia szereg atrybutów warunkowych oraz jeden atrybut odpowiadający klasie danego obiektu. Atrybut decyzyjny jest związany z celem analizy i jego obecność nie podlega dyskusji. Problem znalezienia zbioru znaczącego możemy dalej rozpatrywać tylko dla atrybutów warunkowych. Ogólnie różne zbiory tych atrybutów mogą być porównywane przez ocenianie, jak dobrze reprezentują informację o pojęciu decyzyjnym w stosunku do informacji zawartej pierwotnie w eksperymencie.

W naszym przypadku każdy atrybut warunkowy określa poziom ekspresji odpowiedniego genu. Trzeba zaznaczyć, że wybór ten jest oparty na pewnych założeniach, które mogą w przyszłości okazać się zbyt ostre.

Na wstępie warto przypomnieć, że charakteryzowanie stanu organizmu poprzez opis ekspresji jego genów posiada silną podbudowę biochemiczną. Spośród wielu innych dostępnych parametrów geny uważane są za nośniki kompletnej informacji o białkach produkowanych w organizmie, które z kolei odgrywają kluczową rolę w regulacji i sterowaniu.

Drugim aspektem, jest szczegółowość opisu ekspresji każdego genu. Przyjmuje się, że wystarczająca jest informacja na temat poziomu ekspresji rozumianego przez stężenie kopii danego genu w środowisku komórek organizmu. Stężenie to jest przekładane na ilość produkowanego białka danego typu. Obecnie wszelkie podejmowane kroki dotyczą jedynie rozbudowy informacji na temat poziomu ekspresji np. o oszacowanie błędu pomiaru.

Założenia te uprawniają do ustalenia zbioru atrybutów warunkowych na wszystkie geny genomu danego organizmu. Dalsze ograniczenia zbioru znaczącego wymagają dodatkowej wiedzy na temat badanego zagadnienia. Może ona pochodzić z zewnątrz lub być oparta na wnioskach wysnutych z samej tablicy decyzyjnej. Przykładem pierwszego podejścia może być wykluczenie zbadanych już grup funkcjonalnych genów, albo materiału z chromosomów płci u człowieka. Dalej rozpatrujemy wyłącznie dane wynikowe eksperymentu i zakładamy, że podejście to zostało wykorzystane na etapie planowania przez zespół biologiczny.

W rozdziale skupimy się na metodach wyboru atrybutów poprzez analizę tablicy decyzyjnej. Miejmy jednocześnie na uwadze, że kolejne etapy analizy, które wykorzystują metody odkrywania wiedzy, zmierzają do wykrycia grup genów o szczególnym znaczeniu dla badanej hipotezy. Przeprowadzona przez nas eliminacja atrybutów nieistotnych nie powinna niekorzystnie wpłynąć na wyniki otrzymywane za pomocą tych metod.

Założmy, że pojęcie siły powiązania zbioru atrybutów jest dobrze zdefiniowane. W przypadku idealnym, dla określonego poziomu siły powiązania, znaleziony zbiór znaczący powinien być nadzbiorem dla wszystkich grup atrybutów odpowiednio silnie związanych z badaną hipotezą, które zostałyby wyznaczone bez ograniczania opisu obiektów. W praktyce oczekujemy wyraźnego podobieństwa pomiędzy wynikami otrzymanymi przy pełnym i ograniczonym zestawie atrybutów.

4.2. Znaczenie

Wyznaczenie atrybutów istotnych dla reprezentacji pojęcia decyzyjnego ma zarówno znaczenie informacyjne, jak i praktyczne.

Zbiór atrybutów znaczących zawiera bezpośrednio wiedzę o tym, jakie geny decydują o różnym zachowaniu organizmu. Wiedza stanowi już pewien istotny wkład w badanie zagadnienia. Nawet w przypadku prymitywnych organizmów rozmiary genomu wielokrotnie przewyższają rozmiary wyznaczanych zbiorów genów znaczących.

Częstym zastosowaniem praktycznym jest rozłożenie badania danej hipotezy na kilka etapów. Podstawowym kryterium jest tu koszt i czas przeprowadzenia pojedynczego eksperymentu. Duża ilość materiału genetycznego pociąga za sobą droższą technologię. Etapy prowadzą do znalezienia odpowiednio małego górnego ograniczenia zbioru znaczącego. Badania często rozpoczynają się od niewielu pomiarów stanu organizmu dla całego genomu i ustalenia, które geny są najmniej znaczące. Kolejne iteracje operują na coraz węższym zbiorze atrybutów warunkowych i pozwalają na przeprowadzenie coraz większej liczby pomiarów stanu. Dopiero ostatni eksperyment jest punktem wyjścia do szczegółowej analizy badanej w doświadczeniu hipotezy.

Mała liczba atrybutów warunkowych w tablicy decyzyjnej jest także kluczowa dla efektywności wielu zaawansowanych algorytmów, w szczególności dla licznych strategii odkrywania wiedzy. Wstępna eliminacja części atrybutów warunkowych okazuje się konieczna, jeśli problem charakteryzuje się wysoką wymiarowością.

W przypadku eksperymentów z wykorzystaniem mikromacierzy zakres badanego genomu jest zwykle wiele większy niż liczba przeprowadzonych hybrydyzacji. Z tego względu w wyniku otrzymujemy tablicę decyzyjną o bardzo dużej liczbie atrybutów i stosunkowo małej liczbie obiektów. Różnica ta wynosi najczęściej dwa rzędy wielkości.

4.3. Statystyczna ocena atrybutów

4.3.1. Wprowadzenie

Tablica decyzyjna jest zapisem pojedynczego eksperymentu. Rozważane przez nas uniwersum stanowi wyłącznie próbę losową spośród możliwych stanów organizmów. Do przeprowadzenia analizy wartości atrybutów obiektów próby możemy posłużyć się podstawowym aparatem statystycznym.

Dalej stosowane są pojęcia i oznaczenia zgodne z wprowadzonymi w rozdziale 3.2. Ze względu na przyjęte ogólnie nazewnictwo termin grupa będzie używany wymiennie z klasą.

W tym rozdziale poruszone jest wyłącznie problem oszacowywanie związku pojedynczego atrybutu warunkowego z atrybutem decyzyjnym, pominięte jest zagadnienie oceny podzbiorów atrybutów. Pomimo swojej prostoty oceny tego typu są obecnie podstawowym narzędziem znajdowania genów znaczących.

4.3.2. Jednoczynnikowa analiza wariancji

Jednoczynnikowa analizy wariancji, znana z angielskiego pod nazwą One-way ANOVA (ANalysis Of Variance) jest powszechnie stosowanym testem statystycznym. Na polu badania ekspresji genów po raz pierwszy zastosowali ją Kerr Kathleen i Gary Churchill w 2000 roku [26]. W ogólności może ona być wykorzystana do oszacowania wpływu pojedynczego atrybutu na wartość atrybutu decyzyjnego.

Rozpatrujemy wybrany atrybut warunkowy a_1 i atrybut decyzyjny d .

Podzielmy populację wartości atrybutu a_1 dla wszystkich obiektów uniwersum na p populacji odpowiadających kolejnym wartościom atrybutu decyzyjnego. W ramach każdej populacji wprowadźmy osobną numerację.

Przyjęte oznaczenia i podstawowe wielkości:

- n_i – liczności populacji i -tej
- y_{ij} – j -ta wartość atrybutu w ramach i -tej populacji
- $\bar{y}$ – średnia wszystkich wartości, przy czym $\bar{y} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij}$
- $\bar{y}_i$ – średnia grupowa populacji i -tej klasy, przy czym $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$
- v_i – wariancja populacji i -tej klasy, przy czym $v_i = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$

Istotą analizy wariancji jest oddzielne rozpatrywanie dwóch rodzajów zróżnicowania wartości atrybutu:

- Różnic pomiędzy klasami spowodowanych badanym atrybutem
- Różnic występujących wewnątrz klas spowodowanych błędami doświadczalnymi

Pierwszy składnik, zwany międzygrupowym, mierzy odchylenia średnich klas względem średniej ogólnej i przyjmuje postać:

$$SS_{between} = n \sum_{i=1}^p (\bar{y}_i - \bar{y})^2$$

Drugi składnik, zwany wewnątrzgrupowym, mierzy odchylenia wartości w klasach względem średniej tych dla klas. Przyjmuje on postać:

$$SS_{within} = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Całkowita zmienność atrybutu wyrażona jest przez:

$$SS_{total} = \sum_{i=1}^p \sum_{j=1}^n (y_{ij} - \bar{y})^2$$

Liczby stopni swobody dla poszczególnych źródeł zmienności wynoszą odpowiednio:

$$df_{between} = p - 1$$

$$df_{within} = p(n - 1)$$

$$df_{total} = pn - 1$$

Nieobciążonymi miarami odpowiednich wariancji z próby są średnie kwadraty odchyłeń:

$$MS_{between} = SS_{between} / df_{between}$$

$$MS_{within} = SS_{within} / df_{within}$$

$$MS_{total} = SS_{total} / df_{total}$$

Statystyka F mierzy stopień rozbieżności pomiędzy hipotezą, że wartości średnich grupowych p populacji wartości atrybutów są takie same, a sytuacją rzeczywistą:

$$F_{a_1} = MS_{between} / MS_{within}$$

Statystyka może zostać wykorzystana do oszacowania zależności atrybutu decyzyjnego od badanego atrybutu warunkowego. Im większa wartość F tym wpływ na klasę jest silniejszy. Dla $F=1$ rozważany atrybut nie ma wpływu na klasę.

Warto dodać, że powyższe podejście jest uogólnieniem testu t . Statystyka t pozwala wyłącznie na oszacowanie związku atrybutu warunkowego z atrybutem decyzyjnym przyjmującym dwie wartości. W badaniach nad ekspresją ograniczenie to często nie jest istotne. Metoda znajduje szerokie zastosowanie w licznych eksperymentach [4,50], które dotyczą przykładów z dwóch klas decyzyjnych.

Statystyka t ma postać: $t_{a_1} = (\bar{y}_1 - \bar{y}_2) / \sqrt{\frac{v_1}{n_1} + \frac{v_2}{n_2}}$

Zastosowanie bezpośrednio statystyki t dla każdej możliwej 2-elementowej kombinacji klas jest odradzane z kilku względów:

- Liczba statystyk t do obliczenia rośnie kwadratowo z liczbą klas, konkretnie dla p klas należy przeprowadzić $\binom{p}{2}$ obliczeń statystyki
- Z liczbą obliczonych statystyk rośnie prawdopodobieństwo zajścia błędu pierwszego rodzaju, czyli odrzucenia hipotezy prawdziwej
- Braku uzasadnionej formalnie metody odniesienia do siebie statystyk dla różnych klas

Dla przypadku dwóch klas decyzyjnych istnieje prosty związek pomiędzy poruszonymi statystykami, który przyjmuje postać $F_{a_1} = t_{a_1}^2$.

Indeksy przy statystykach F i t wskazują na atrybut warunkowy tablicy decyzyjnej, dla którego zostały wyznaczone. Dla przejrzystości zapisu indeks taki został pominięty przy opisywaniu wielkości cząstkowych.

Przykład

Rozważmy tablicę decyzyjną DT_j . Dla ocenienia znaczenia atrybutów zostanie wykorzystana jednoczynnikowa analiza wariancji. Zauważmy, że przypadek dotyczy dwóch klas decyzyjnych, więc do identycznych wyników końcowych doprowadzi również obliczenie statystyki t . Ze względu na przejrzystość schematu przykład prześledzi obliczenia dla metody analizy wariancji.

Tabela 4.3.2-1 zawiera wyniki wspólne dla wszystkich atrybutów oraz cząstkowe dla genu G_1 . Dla pozostałych genów obliczenia są analogiczne. Tabela 4.3.2-2 zawiera istotniejsze wyniki cząstkowe oraz ostateczne dla wszystkich atrybutów.

y_{ij}	1	2	3	4
1	1,01	1,74	1,52	1,02
2	3,17	4,12	8,17	9,65

I	n_i	$\bar{y}_i$	v_i
1	4	1,32	32,00
2	4	6,28	52,25

$\bar{y}$	3,80
$SS_{between}$	49,10
SS_{within}	255,75
SS_{total}	157,55

Tab. 4.3.2-1. Wyniki cząstkowe dla atrybutu G_1 ($a_2=d$)

Atrybut	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
$df_{between}$	1	1	1	1	1	1	1	1
df_{within}	6	6	6	6	6	6	6	6
df_{total}	7	7	7	7	7	7	7	7
$MS_{between}$	49,10	1,94	27,86	18,24	21,78	20,83	0,014	0,446
MS_{within}	42,63	1,38	21,95	19,64	17,39	33,55	7,656	7,034
MS_{total}	22,51	0,63	10,85	11,62	8,68	22,80	6,558	5,901
t	1,15	1,40	1,27	0,93	1,25	0,62	0,001	0,063
F	1,07	1,18	1,13	0,96	1,12	0,79	0,043	0,251

Tab. 4.3.2-2. Wyniki ostateczne dla atrybutów warunkowych

Wyznaczone wartości statystyki F wskazują, że atrybuty mogą zostać uporządkowane zgodnie z wielkością związku z hipotezą docelową w następujący sposób: $G_2, G_3, G_5, G_1, G_4, G_6, G_8, G_7$.

4.3.3. Statystyka χ^2

Statystyka χ^2 jest popularną metodą stosowaną do oceny prawdopodobieństwa, że dwie zmienne dyskretne są od siebie niezależne. Statystyka mierzy różnicę między faktycznym rozkładem poszczególnych par wartości dwóch atrybutów, a ich rozkładem oczekiwanym przy założeniu niezależności atrybutów. Im bardziej różnią się oba rozkłady, tym większa jest wartość statystyki.

W przypadku ocenienia zależności warunkowego atrybutu ciągłego i dyskretnego atrybutu decyzyjnego niezbędny jest etap wstępnej dyskretyzacji. Tematyka dyskretyzacji jest przedmiotem rozdziału 5. Tutaj będziemy operować na wybranej zdyskretyzowanej tablicy decyzyjnej.

Założmy, że badamy zależność dwóch atrybutów dyskretnych a_1 i a_2 , takie że $a_1 : U \rightarrow V_1$ i $a_2 : U \rightarrow V_2$.

Oznaczmy przez:

- W_i – zbiór wartości atrybutu a_i , przy czym $W_i = a_i(U)$, $W_i = \{w_i^1, \dots, w_i^{k_i}\}$, dla $k_i = |W_i|$
- f_i^k – liczbę wystąpień i -tej wartości atrybutu k
- f_{ij} – liczbę wystąpień i -tej wartości atrybutu 1 i j -tej wartości atrybutu 2 jednocześnie
- e_{ij} – oczekiwaną liczbę wystąpień i -tej wartości atrybutu 1 i j -tej wartości atrybutu 2 jednocześnie przy założeniu niezależności atrybutów

Statystyka ma postać:

$$\chi^2_{a_1, a_2} = \sum_{i \in W_1} \sum_{j \in W_2} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Wielkości wyznaczamy przez:

$f_i^k = \{u \in U : a_k(u) = w_k^i\} $	$f_{ij} = \{u \in U : a_1(u) = w_1^i \wedge a_2(u) = w_2^j\} $	$e_{ij} = \frac{f_i^1 f_j^2}{m}$
--	---	----------------------------------

Statystyka χ^2 jest zmienną losową mającą przybliżony rozkład χ^2 o $(|W_1|-1)(|W_2|-1)$ stopniach swobody. Przy zadanym poziomie istotności h można wyznaczyć na podstawie tablic wartość progową statystyki taką, że prawdopodobieństwo jej uzyskania lub przekroczenia dla niezależnych atrybutów wynosi h .

Można również posłużyć się znormalizowaną postacią statystyki χ^2 w postaci współczynnika Cramera:

$$V_{a_1, a_2} = \sqrt{\frac{\chi^2}{m(t-1)}}, \text{ przy czym } t = \min(|W_1|, |W_2|)$$

Postać ta jest związana z maksymalnym poziomem istotności, a otrzymywana jest przy założeniu, że statystyka χ^2 ma wartość 0 dla równości rozkładu faktycznego i oczekiwanego częstości występowania wartości atrybutów.

Powyższa statystyka szacuje zależność pomiędzy dwoma atrybutami. W szczególności może ona być użyta do zbadania zależności wybranego atrybutu warunkowego z atrybutem decyzyjnym. Postępowanie to można zastosować do wszystkich atrybutów tablicy decyzyjnej. Otrzymujemy wtedy ogólną informację na temat wpływu każdego atrybutu warunkowego na decyzje.

Indeksy przy statystykach χ^2 i V wskazują na parę atrybutów tablicy decyzyjnej, których związek jest szacowany. Dla przejrzystości zapisu indeks taki został pominięty przy opisywaniu wielkości częściowych.

Przykład

Znaczenia atrybutów zostanie ocenione za pomocą znormalizowanej statystyki χ^2 . Wyznaczenie statystyki wymaga wstępnej dyskretyzacji tablicy decyzyjnej. Z tego względu rozważymy tablicę decyzyjną DT_3 , która została otrzymana w wyniku dyskretyzacji tablicy DT_1 z równą szerokością przedziałów. Wykorzystana metoda dyskretyzacji została szerzej przedstawiona w rozdziale 5.3.1. Po zakończeniu procesu mamy $W_1 = \{0, 1\}$ i $W_2 = \{Z, R\}$. Dla przejrzystości przykładu tablica ta została zamieszczona również w tabeli 4.3.3-1.

Tabela 4.3.3-2 zawiera wyniki częściowe dla genu G_1 , dla pozostałych genów obliczenia są analogiczne. Tabela 4.3.3-3 zawiera wartości statystyki, liczby stopni swobody i współczynnika Cramera dla wszystkich genów. Wyznaczone oceny zostaną wykorzystane do podsumowania znaczenia genów w rozdziale 8.

	G_1	G_2	G_3	G_4	G_5	d
H_1	0	0	0	1	1	Z
H_2	0	0	0	1	1	Z
H_3	0	0	1	0	0	Z
H_4	0	0	1	1	1	Z
H_5	0	1	0	0	0	R
H_6	0	1	0	0	0	R
H_7	1	1	0	1	0	R
H_8	1	1	0	0	0	R

Tab. 4.3.3-1. Zdyskretyzowana tablica decyzyjna DT_3 przy zachowaniu równej szerokości przedziałów

f_i^k	k=1	k=2	f_{ij}	j=1	j=2	e_{ij}	j=1	j=2
i=1	6	4	i=1	4	2	i=1	3,0	3,0
i=2	2	4	i=2	0	2	i=2	1,0	1,0

Tab. 4.3.3-2. Wyniki cząstkowe dla atrybutu G_1 ($a_2=d$)

Atrybut	G_1	G_2	G_3	G_4	G_5
t	2	2	2	2	2
χ^2	2,67	8,00	2,67	2,00	4,80
V	0,58	1,00	0,58	0,50	0,77

Tab. 4.3.3-3. Wyniki ostateczne dla atrybutów warunkowych

Wyznaczone wartości statystyki χ^2 wskazują, że atrybuty mogą zostać uporządkowane zgodnie z wielkością związku z hipotezą docelową w następujący sposób: G_2, G_5, G_3, G_1, G_4 . Widać, że porządek ten różni się od porządku pomiędzy odpowiednimi atrybutami ciągłymi wskazanego przez statystykę F .

4.4. Pojęcie zbioru znaczącego

4.4.1. Ocena znaczenia zbioru atrybutów

Rozpatrzmy tablicę decyzyjną $(U, A, C, \{d\})$. Przestrzeń możliwych zbiorów atrybutów równa się 2^C i posiada rozmiar wynoszący $2^{|C|}$. Akceptujemy tu zbiór pusty, który może być poprawnym rozwiązaniem dla pewnych zdegenerowanych przypadków tablicy decyzyjnej, np. tablica o jednoelementowym uniwersum.

Zauważmy na wstępie, że zbioru atrybutów znaczących nie można budować na podstawie badania znaczenia każdego atrybutu z osobna. Stąd, wybranie pewnej liczby najbardziej znaczących atrybutów zgodnie z pewną miarą i potraktowanie ich jako zbioru, który najlepiej będzie reprezentował pojęcie decyzyjne jest nieuzasadnione. Poprawne podejście powinno zbadać, jaki związek z klasą mają kolejne podzbiory atrybutów.

Dla analizy istotności zbiorów wprowadzimy funkcję *setval* wyrażającą siłę związku danego zbioru atrybutów z hipotezą badaną. Ważny jest fakt, że miara dotyczy hipotezy reprezentowanej przez cały rozważany zbiór atrybutów.

W wielu przypadkach wyznaczenie znaczenia pojedynczego atrybutu jest przeprowadzane odmiennie niż wyznaczanie znaczenia podzbioru atrybutów. Uzasadnia to zdefiniowanie funkcji oceny za pomocą dwóch różnych funkcji.

Definicja 4.4.1-1

Funkcja oceny znaczenia zbioru atrybutów oznacza funkcję $setval: 2^C \rightarrow \langle 0, 1 \rangle$:

$$setval(X) = \frac{1}{M_{setval}} setval'(X)$$

gdzie:

- $attrval: C \rightarrow R^+ \cup \{0\}$
- $attrsetval: 2^C \rightarrow R^+ \cup \{0\}$
- $setval'(X) = \begin{cases} 0 & X = \emptyset \\ attrval(x) & X = \{x\} \\ attrsetval(X) & |X| > 1 \end{cases}$
- $M_{setval} = \max_{X \in 2^C} (setval'(X))$

W dalszej części rozdziału wartość danego zbioru atrybutów będzie utożsamiana z wartością funkcji *setval* dla tego zbioru. W celu ograniczenia zbioru wartości do przedziału $\langle 0, 1 \rangle$ funkcja została znormalizowana względem wartości zbioru atrybutów najbardziej związanego z hipotezą. Nie wprowadza to istotnego ograniczenia użyteczności definicji, gdyż w rozpatrywanych dalej zastosowaniach konieczne jest wyznaczenie wartości funkcji dla całej dziedziny. Zaznaczmy, że wybór zbioru $\langle 0, 1 \rangle$ na przeciwdziedzinę funkcji jest arbitralny, zamiast niego można zaproponować dowolny inny zbiór uporządkowany liniowo. Wybór uzasadniony jest łatwością operowania wartościami znormalizowanymi.

Funkcja *attrval* wyraża związek pojedynczego atrybutu z badaną hipotezą. Znaczenie podzbioru atrybutów opisywane jest przez funkcję *attrsetval*. Podkreślmy, że powyższa definicja funkcji *setval* wynika często ze stosowania różnych strategii dla oceny pojedynczych atrybutów i ich zbiorów. Pomijając to założenie funkcja *attrval* mogłaby być wyrażona w szczególności za pomocą funkcji *attrsetval*, co prowadziłoby do równoważności funkcji *setval* i znormalizowanej funkcji *attrsetval*. Zauważmy, że wartości wskazywane przez funkcję *attrsetval* dla zbioru pustego i jednoelementowych zbiorów atrybutów są nieistotne dla ostatecznej wartości znaczenia danego podzbioru obliczonej przy pomocy funkcji *setval*. Wynika stąd, że do zdefiniowania *setval* wystarczy funkcja obcięta $attrsetval|_{\{X \in 2^C: |X| > 1\}}$.

Zdefiniowanie funkcji *attrval* i *attrsetval* tak, aby odzwierciedlały związek odpowiednio pojedynczego atrybutu i ich zbioru zależy od charakteru jej atrybutów. W pracy przedstawimy ogólne podejście dla tablicy decyzyjnej zawierającej dowolne atrybuty, a następnie rozważymy dwa szczególne przypadki tablicy z wszystkimi atrybutami warunkowymi ciągłymi i dyskretnymi. Dla przypadków szczególnych proponujemy podejścia alternatywne w stosunku do podejścia ogólnego.

4.4.1.1. Przypadek atrybutów dowolnych

Do oceny znaczenia zbioru atrybutów można wykorzystać algorytmy uczenia pod nadzorem. W zależności od charakteru atrybutów należy dobrać taki klasyfikator, aby przekształcenie atrybutów do akceptowanej przez niego postaci wprowadzało minimalną stratę informacji.

Ocena zbioru jest równa średniej dokładności wybranego klasyfikatora. Ponieważ tablica zawiera wszystkie dostępne nam obiekty, więc należy zdecydować się na odpowiednią politykę podziału uniwersum na zbiór trenujący i testowy. Przykładowo można zastosować walidację krzyżową.

Ocena pojedynczego atrybutu jest równa ocenie zbioru jednoelementowego zawierającego ten atrybut.

4.4.1.2. Przypadek atrybutów ciągłych

Do badania znaczenia pojedynczego atrybutu w wielu przypadkach wykorzystywane jest podejście statystyczne. W szczególności dla znajdowania genów znaczących najczęściej stosowana jest przedstawiona w rozdziale 4.3.2 metoda analizy wariancji.

Dla wygody wprowadzimy funkcję $attrval_F: C \rightarrow R^+ \cup \{0\}$, zdefiniowaną dla dowolnego $a \in C$ następująco: $attrval_F(a) = \frac{1}{F_{\max}} F_a$, przy czym $F_{\max} = \max_{a \in C} (F_a)$.

W pracy nie jest dyskutowana możliwość rozszerzenia tej miary na podzbiory atrybutów.

4.4.1.3. Przypadek atrybutów dyskretnych

Podobnie jak w przypadku atrybutów ciągłych szacowanie znaczenia atrybutów dyskretnych również często odwołuje się do metod statystycznych. W rozdziale 4.3.3 do tego celu zaproponowaliśmy statystykę χ^2 .

Dla wygody wprowadzimy funkcję $attrval_\chi: C \rightarrow R^+ \cup \{0\}$, zdefiniowaną dla dowolnego $a \in C$ następująco: $attrval_\chi(a) = V_{a,d}$.

Funkcja $attrsetval$ powinna być tak zdefiniowana, aby jej wartości można było porównywać z ocenami pojedynczych atrybutów. Do osiągnięcia tego celu posłużymy się modyfikacją tablicy decyzyjnej, która nie wprowadzi straty informacji w stosunku do postaci pierwotnej.

Definicja 4.4.1.3-1

Niech dana będzie tablica decyzyjna (U, A, C, D) .

Połączenie M atrybutów warunkowych $a_{i_1}, \dots, a_{i_M}$ polega na wyznaczeniu tablicy decyzyjnej $(U, A', C', \{d\}) = (U, A \cup \{a'\} \setminus \{a_{i_1}, \dots, a_{i_M}\}, C \cup \{a'\} \setminus \{a_{i_1}, \dots, a_{i_M}\}, \{d\})$, gdzie $a': U \rightarrow V_{i_1} \times \dots \times V_{i_M}$.

Połączenie atrybutów polega na utworzeniu na podstawie atrybutów $a_{i_1}, \dots, a_{i_M}$ nowego atrybutu i zastąpieniu tych atrybutów atrybutem a' . Otrzymana tablica decyzyjna ma mniejszą liczbę atrybutów warunkowych, dokładnie $|C'| = |C| - M + 1$. Dokładniejsze omówienie znajduje się w pracy [10].

Dla wygody zdefiniujemy funkcję pomocniczą $merge: 2^C \rightarrow C'$, gdzie C' jest zbiorem możliwych atrybutów połączonych dla zbioru atrybutów pierwotnych C . Funkcja $merge$ przyporządkowuje danemu zbiorowi M atrybutów nowy atrybut utworzony w wyniku ich połączenia.

Wykorzystując powyższe pojęcia możemy zaproponować definicję funkcji *attrsetval* w postaci:

$$\text{attrsetval}(X) = \begin{cases} 0 & X = \emptyset \\ \text{attrval}_\chi(x) & X = \{x\} \\ \text{attrval}_\chi(\text{merge}(X)) & |X| > 1 \end{cases}$$

Przykład

Rozpatrzmy tablicę decyzyjną DT_3 o wartościach dyskretnych. Tablica ta została wykorzystana do wyznaczenia statystyki χ^2 dla wszystkich atrybutów i jest przedstawiona w tabeli 4.3.3-1. Dokonajmy połączenia 2 atrybutów warunkowych G_3, G_5 .

	G_1	G_2	G_4	G_{35}	d
H_1	0	0	1	(0,1)	Z
H_2	0	0	1	(0,1)	Z
H_3	0	0	0	(1,0)	Z
H_4	0	0	1	(1,1)	Z
H_5	0	1	0	(0,0)	R
H_6	0	1	0	(0,0)	R
H_7	1	1	1	(0,0)	R
H_8	1	1	0	(0,0)	R

Tab. 4.4.1-1. Tablica decyzyjna po połączeniu 2 atrybutów warunkowych: G_3, G_5

Dodajmy, że podobnego spojrzenia na problem dostarczają prace [23,24]. Autor tych prac porównuje podziały zbioru obiektów tablicy decyzyjnej wprowadzane odpowiednio przez oceniany zbiór atrybutów i przez atrybut decyzyjny.

4.4.2. Ocena znaczenia zbioru atrybutów ze względu na podzbiory

W poprzednim rozdziale zdefiniowaliśmy funkcję *setval*, która opisuje związek danego zbioru atrybutów z klasą decyzyjną. Podkreślmy, że ocena ta dotyczy związku hipotezy reprezentowanej dokładnie tym zbiorem atrybutów.

Etap wyznaczenia zbioru znaczącego jest wstępem do zastosowania metod bardziej zaawansowanych. Metody te mogą odwoływać się do różnych podzbiorów wybranego zbioru atrybutów. Z tego względu jesteśmy zainteresowani znalezieniem zbioru, który będzie zawierał wszystkie podzbiory dostatecznie silnie związane z badaną hipotezą. W celu formalnego sprecyzowania pojęcia zbioru znaczącego proponujemy dwa podejścia.

W pierwszym podejściu zdefiniujemy zbiór znaczący w taki sposób, aby zawierał wszystkie zbiory o wartości nie gorszej niż określony próg.

Definicja 4.4.2-1

Niech dana będzie tablica decyzyjna $(U, A, C, \{d\})$ i liczba rzeczywista $\lambda \in \langle 0, 1 \rangle$.

Zbiór **λ -rankingowo znaczący** oznacza zbiór $\bigcup_{X \in F} X$ dla $F(\lambda) = \{X \in 2^C : \text{setval}(X) \geq \lambda\}$.

Zauważmy, że zgodnie z definicją przyjęcie odpowiedniego progu λ może doprowadzić do eliminacji pewnego atrybutu a , wtedy i tylko wtedy, gdy nie znajduje się on w żadnym podzbiorze X , którego ocena znaczenia jest co najmniej λ .

W drugim podejściu w celu badania istotności danego podzbioru atrybutów dla hipotezy wprowadzimy funkcję oceny $val: 2^C \rightarrow \langle 0, 1 \rangle$.

Dla danego argumentu X wartość szukanej funkcji powinna uwzględniać powiązanie dowolnego podzbioru X z pojęciem decyzyjnym. Wymaganie to sprecyzujemy w postaci założenia dla funkcji val .

Założenie 1

Funkcja val jest niemalejąca względem częściowo porządkującej relacji inkluzji zbiorów w przestrzeni 2^C : $\forall (X, Y \in 2^C): X \subset Y \Rightarrow val(X) \leq val(Y)$.

Zbiór pusty jest elementem najmniejszym dla relacji inkluzji zbiorów, dlatego odpowiada mu najmniejsza wartość funkcji val . Dla wygody wartość ta ustalona została na 0, co wyrażone jest w kolejnym założeniu.

Założenie 2

$val(\emptyset) = 0$

Ponieważ zbiór C jest elementem największym dla relacji inkluzji zbiorów, więc funkcja val jest ograniczona z góry przez swoją wartość maksymalną przyjmowaną dla pełnego zbioru atrybutów warunkowych C : $\sup_{X \in 2^C} (val(X)) = \max_{X \in 2^C} (val(X)) = val(C)$. Ponadto, wartość ta może

być wykorzystana jako odniesienie dla wartości pozostałych i pozwolić na stworzenie względnej miary, która byłaby niezależna od zbioru C .

Istnieje nieskończenie wiele funkcji spełniających powyższe dwa wymagania. W pracy ograniczymy się do podania dwóch przykładów. Obie propozycje wykorzystują wprowadzoną wcześniej funkcję $setval: 2^C \rightarrow \mathbb{R}^+ \cup \{0\}$. Przypomnijmy, że wartość funkcji $setval$ dla danego zbioru atrybutów X określa siłę związku pojęcia reprezentowanego dokładnie przez zbiór X z pojęciem decyzyjnym.

$$val_1(X) = \frac{1}{val_1'(C)} val_1'(X) \quad \text{gdzie} \quad val_1'(X) = \begin{cases} 0 & X = \emptyset \\ \max_{Y \in 2^X} (setval(Y)) & X \neq \emptyset \end{cases}$$

$$val_2(X) = \frac{1}{val_2'(C)} val_2'(X) \quad \text{gdzie} \quad val_2'(X) = \begin{cases} 0 & X = \emptyset \\ \sum_{Y \in 2^X} \left(\frac{|C|}{|Y|} setval(Y) \right) & X \neq \emptyset \end{cases}$$

Funkcja val_1 ocenia wartość zbioru X wybierając największą z ocen wszystkich podzbiorów X . Może ona być zastosowana do sytuacji, w której poszukujemy najmniejszego zbioru zawierającego podzbiór najsilniej związany z badaną hipotezą.

Z kolei funkcja val_2 ocenia wartość zbioru X sumując oceny wszystkich podzbiorów X promując dodatkowo zbiory krótkie. W dalszej analizie skupiać się będziemy na badaniu różnych podzbiorów atrybutów i potrzebujemy metody, która wyłoni zbiór o podzbiórach

silnie związanych z hipotezą. Z tego względu do opisu zbioru znaczącego wykorzystamy funkcję val_2 .

Obie funkcje zostały znormalizowane za pomocą swojej wartości dla zbioru C . W ten sposób otrzymujemy ograniczenie zbioru wartości tych funkcji do przedziału $\langle 0, 1 \rangle$.

Powyższe rozważania prowadzą do innego sformułowania pojęcia zbioru znaczącego.

Definicja 4.4.2-2

Niech dana będzie tablica decyzyjna $(U, A, C, \{d\})$ i liczba rzeczywista $\lambda \in \langle 0, 1 \rangle$.

Zbiór $X \subset C$ nazywamy **λ -sumacyjnie znaczącym**, jeśli jest najmniejszym zbiorem takim, że $val_2(X) \geq \lambda$ i nie istnieje zbiór Y taki, że $X \subsetneq Y$ i $val_2(Y) \geq \lambda$.

Parametr λ może być traktowany, jako próg istotności zbioru dla badanej hipotezy. Ze względu na znormalizowaną postać funkcji oceny wygodnie jest wyrażać go w procentach.

Zastanówmy się teraz, jak złożonym problemem jest wyznaczenie zbioru znaczącego rozumianego zgodnie definicją 4.4.1-2 i 4.4.1-3.

W przypadku obu podejść do znalezienia zbioru znaczącego dla danej wartości λ niezbędne jest obliczenie wartości funkcji $setval$ dla całej jej dziedziny. Potraktujmy złożoność tego podproblemu jako dolne ograniczenie na złożoność wyznaczenia całej funkcji val_2 .

W rozważaniach przyjmijmy obliczenie funkcji $attrval$ dla konkretnego atrybutu jako operację jednostkową. Wyznaczenie zbioru wartości $setval$ wymaga rozważenia wszystkich elementów przestrzeni 2^C , w tym przeprowadzenia połączeń atrybutów dla zbiorów o liczbie elementów przekraczającej 1. Obserwacja ta eliminuje podejście wyczerpujące, jako praktyczną drogę do oceniania wartości zbiorów atrybutów. Z tego względu zmuszeni jesteśmy do wyboru metody przybliżonej.

W celu wyznaczenia zbioru λ -rankingowo znaczącego proponujemy rozważenie jedynie wybranej rodziny podzbiorów C . Wybór rodziny uzasadniony jest względami efektywnościowymi i będzie brał pod uwagę kolejno coraz liczniejsze kombinacje atrybutów z X .

Wprowadźmy następujące oznaczenie zbioru kombinacji: $K_M(X) = \{Y \subset X : |Y| \leq M\}$

Definicja 4.4.1-3

Niech dana będzie tablica decyzyjna $(U, A, C, \{d\})$, liczba rzeczywista $\lambda \in \langle 0, 1 \rangle$, liczba naturalna $M \in \{0, \dots, n\}$ oraz funkcja $setval: 2^C \rightarrow R^+ \cup \{0\}$.

Zbiór **λ, M -rankingowo znaczący** oznacza $\bigcup_{X \in F_M} X$ dla $F_M(\lambda) = \{X \in K_M(C) : setval(X) \geq \lambda\}$.

Analogiczny zabieg można zastosować w drugim podejściu przeddefiniowując funkcję val_2 tak, aby rozważała wyłącznie określoną rodzinę podzbiorów argumentu X .

Definicja 4.4.1-4

Niech dana będzie tablica decyzyjna $(U, A, C, \{d\})$, liczba rzeczywista $\lambda > 0$, liczba naturalna $M \in \{0, \dots, n\}$ oraz funkcja $setval: 2^C \rightarrow R^+ \cup \{0\}$.

Zbiór $X \subset C$ nazywamy **λ, M -sumacyjnie znaczącym**, jeśli jest najmniejszym zbiorem takim, że $val_2'(X) \geq \lambda$ i nie istnieje zbiór Y taki, że $X \subsetneq Y$ i $val_2'(Y) \geq \lambda$. Przy czym $val_2': 2^C \rightarrow R^+ \cup \{0\}$ jest zdefiniowana następująco:

$$val(X) = \frac{1}{val'(C)} val'(X) \quad \text{gdzie} \quad val'(X) = \begin{cases} 0 & X = \emptyset \\ \sum_{Y \in K_M(X)} \left(\frac{|C|}{|Y|} setval(Y) \right) & X \neq \emptyset \end{cases}$$

W obu podejściach parametr M można interpretować jako dokładność oceny zbioru X , a λ jako miarę istotności zbioru.

Zastosowanie zbioru $K_M(X)$ do wyznaczenia przybliżonej wartości zbiorów znaczących jest uzasadnione łatwą kontrolą złożoności obliczeń dla określonych M . Możliwe są inne strategie doboru elementów przestrzeni 2^C , w szczególności heurystyki rozważające podzbiory różnego rozmiaru i opierające swoje działanie na obliczeniach wykonanych już dla innych zbiorów.

Podsumowując, problem wyznaczenia zbioru znaczącego możemy rozumieć formalnie jako problem wyznaczenia zbioru λ, M -rankingowo znaczącego lub λ, M -sumacyjnie znaczącego dla określonego progu λ i dokładności oszacowania M .

Dla zachowania jednoznaczności zapisu do rozdzielania parametrów często będziemy stosować średnik w miejsce przecinka.

Przykład

Rozważmy tablicę decyzyjną DT_1 . Wyznamy zbiory znaczące dla $M=1$ opierając się na wartościach funkcji $attrval_F$ obliczonych na podstawie wartości statystyki F zawartych w tabeli 4.3.2-2. Dla przejrzystości wielkości te zostały przytoczone w tablicy 4.4.2-2. Wartości λ zostały dobrane eksperymentalnie na podstawie rozkładu wartości statystyki dla atrybutów. Dokładniej problem doboru parametrów zbioru znaczącego poruszony jest w rozdziale 4.4.4.

Atrybut	G ₁	G ₂	G ₃	G ₄	G ₅	G ₆	G ₇	G ₈
F	1,07	1,18	1,13	0,96	1,12	0,79	0,043	0,251
$attrval_F$	0,82	1,00	0,90	0,66	0,89	0,44	0,00	0,05

Tab. 4.4.2-1. Wartości statystyki F i funkcji $attrval_F$ dla atrybutów tablicy TD_1

Zbiór atrybutów warunkowych $\{G_1, G_2, G_3, G_4, G_5\}$ jest $0,5; 1$ -rankingowo znaczący, przy czym $F_1(0,5) = \{\{G_1\}, \{G_2\}, \{G_3\}, \{G_4\}, \{G_5\}\}$. Pozostałe atrybuty miały istotnie niższą wartość statystyki F , stąd ich eliminacja wydaje się uzasadniona.

Zbiór atrybutów warunkowych $\{G_1, G_2, G_3, G_4, G_5\}$ jest $0,92; 1$ -sumacyjnie znaczący, przy czym $val(\{G_1, G_2, G_3, G_4, G_5\}) = 4,72$. Współczynnik λ został tak dobrany, aby istniał zbiór znaczący różny od C .

Można pokazać na drodze eksperymentu, że oba podejścia wskazują te same atrybuty dla większego zakresu odpowiednich współczynników λ .

Tabela 4.4.1-2 zawiera zredukowaną tablicę decyzyjną zawierającą atrybuty warunkowe ze wskazanego przez oba podejścia zbioru znaczącego. Otrzymana tablica zostanie wykorzystana do dalszej analizy i oznaczona przez DT_2 .

	G ₁	G ₂	G ₃	G ₄	G ₅	d
H ₁	1,01	0,21	2,41	5,51	6,44	Z
H ₂	1,74	0,11	3,12	8,34	6,26	Z
H ₃	1,52	0,12	5,22	3,32	4,12	Z
H ₄	1,02	0,03	6,33	6,02	8,02	Z
H ₅	3,17	1,00	0,37	1,17	3,02	R
H ₆	4,12	1,53	0,36	3,12	3,18	R
H ₇	8,17	0,99	0,91	5,17	2,11	R
H ₈	9,65	0,89	0,51	1,65	3,33	R

Tab. 4.4.2-2. Tablica decyzyjna DT_2 o znaczącym zbiorze atrybutów

4.4.3. Wyznaczanie zbiorów znaczących

Rozpatrzmy problem wyznaczania zbioru λ, M -rankingowo znaczącego. Jak zostało zaznaczone w poprzednim rozdziale potrzebne jest obliczenie wartości funkcji $setval$ dla elementów $K_M(C)$. Na podstawie informacji o wartości zbiorów atrybutów możemy zgodnie z definicją 4.4.1-4 skonstruować rodzinę F_M , a następnie wyznaczyć szukany zbiór znaczący przez zsumowanie jej elementów. Procedura została przedstawiona poniżej, przy czym nie posłużono się w niej bezpośrednio krokiem budowania rodziny F_M .

```

funkcja calculate-ranking-significant-set ( $\lambda, M$ )
 $S \leftarrow \emptyset$ 
wyznacz rodzinę  $K_M(C)$ 
dla każdego  $X \in K_M(C)$ 
    jeśli  $setval(X) \geq \lambda$  to
         $S \leftarrow S \cup X$ 
zwróć  $S$ 

```

Koszt wykonania zależy od wielkości rodziny $K_M(C)$ i liczności jej elementów. W pierwszej kolejności należy skonstruować rodzinę $K_M(C)$, a następnie wyliczane są wartości funkcji $setval$ dla jej elementów. Wymaga to przeprowadzenia $\sum_{k=2}^M \binom{n}{k}$ połączeń oraz wyliczenia

$\sum_{k=0}^M \binom{n}{k}$ wartości funkcji $attrval$ dla rozpatrywanych atrybutów. Należy dodatkowo uwzględnić koszt wykonania każdego połączenia, który zależy od charakteru atrybutów i zastosowanego algorytmu.

Zajmijmy się teraz wyznaczeniem zbioru λ, M -sumacyjnie znaczącego. Zgodnie z definicją 4.4.1-5 należy znaleźć nie tylko zbiór X , dla którego $val_2'(X) \geq \lambda$, ale także zbadać inne zbiory posiadające tę cechę.

Bez dodatkowej wiedzy na temat właściwości zbiorów λ, M -sumacyjnie znaczących jedyną metodą na wyznaczenie szukanego zbioru pozostaje obliczenie wartości funkcji val_2' dla jej dziedziny i identyfikacja zbioru o odpowiednich cechach w całej rodzinie. Zauważmy, że aby nie powtarzać obliczania wartości poszczególnych zbiorów atrybutów warto na wstępie wyznaczyć funkcję $setval$ dla całej dziedziny. Powyższe rozumowanie zostało sprecyzowane w procedurze:

funkcja calculate-summing-significant-set(λ, M)
 $S \leftarrow \emptyset$
 wyznacz funkcję setval dla $K_M(C)$
 wyznacz funkcję val_2' dla $K_M(C)$
 $F \leftarrow \{X \in K_M(C) : val_2'(X) \geq \lambda\}$
 $\{X_1, \dots, X_{|F|}\} \leftarrow$ zbiór elementów F uporządkowany niemalejąco wg $val_2'(X_i)$
dla każdego i od 1 do $|F|$
 jeśli istnieje $X_j : X_j \subset X_i$ **to**
 $X_i \leftarrow X_i \cup X_j$
zwróć S

Nie będziemy się tu zajmować poszukiwaniem efektywniejszych metod wyznaczania zbioru λ, M -sumacyjnie znaczącego dla dowolnego M . Dodamy tylko, że wyznaczanie istotności dla większych zbiorów w kolejnych krokach głównej pętli może być przeprowadzone dopiero, gdy będzie to konieczne. Pesymistyczny przypadek zachodzi, gdy zbiorem λ, M -sumacyjnie znaczącym jest C .

4.4.4. Wyznaczanie zbioru $\lambda, 1$ -sumacyjnie znaczącego

W rozdziale zajmiemy się szczególnym przypadkiem wyznaczania zbioru $\lambda, 1$ -sumacyjnie znaczącego i podamy dla niego efektywny algorytm.

Przypomnijmy, że dla $M=1$ przy obliczeniu wartości funkcji $val_2'(X)$ musimy przeanalizować rodzinę $K_1(X)$ złożoną z jednoelementowych zbiorów, z których każdy zawiera inny atrybut spośród elementów C . Duży wpływ na efektywność algorytmu proponowanego rozwiązania rozważanego przypadku mają następujące obserwacje:

- $|K_1(X)| = |C| + 1$
- $\forall (X, Y \in K_1(C)) : X \cap Y = \emptyset$
- Wyznaczenie funkcji setval wymaga wyłącznie wyznaczenia funkcji attrval
- Najmniejszy względem rozmiaru możliwy zbiór $\lambda, 1$ -sumacyjnie znaczący X ($|X|=k$) zawiera k najistotniejszych elementów zbioru C ocenionych zgodnie z funkcją attrval

funkcja calculate-1-summing-significant-set(λ)
 wyznacz funkcję setval dla $K_1(C)$
jeśli setval($\emptyset$) $\geq \lambda$ **to**
 zwróć $\emptyset$
 wyznacz $val_2'(C)$ dla celów normalizacyjnych
 $S \leftarrow \emptyset$
 $F \leftarrow$ zbiór $\{X_1, \dots, X_{|C|+1}\}$ uporządkowany nierosnąco wg setval(X_i)
dla każdego i od 1 do $|F|$
 $S \leftarrow S \cup X_i$
 jeśli $val_2'(S) \geq \lambda$ **to**
 przerwij
jeśli $val_2'\left(C \setminus \arg \min_{X \in S} (setval(X))\right) \geq \lambda$ **to**
 $S \leftarrow C$
zwróć S

Zaprezentowany algorytm jest klasy $O(n \log n)$, co wynika z operacji sortowania rodziny F . Pozostałe kroki mają koszt liniowy. Wyznacza on w pierwszej kolejności najmniejszy zbiór S ,

dla którego $setval(S) \geq \lambda$. Następnie buduje możliwie najwartościowszy zgodnie z funkcją val_2' zbiór elementów, który nie jest nadzbiorem S . Oznaczmy go przez T . Jeśli $val_2'(T) \geq \lambda$, to oznacza, że zbiór S nie jest zbiorem λ, l -sumacyjnie znaczącym i odpowiedzią jest C .

4.4.5. Dobór parametrów zbiorów znaczących

Oddzielnym problemem jest odpowiedni dobór parametrów λ, M przy wyznaczaniu zbioru znaczącego dla konkretnej tablicy decyzyjnej. Zagadnienie to nie jest przedmiotem pracy, więc rozdział ten tylko zarysuje najistotniejsze kwestie.

Dokładność szacowania wartości zbioru znaczącego jest dobierana w zależności od rozmiaru analizowanej tablicy decyzyjnej. Zauważmy, że badanie wszystkich kombinacji $K_M(C)$ dla danego M jest bardzo kosztowne. Ze względu na złożoność czasową wartość M należy ograniczyć się do 2 lub 3. Przy dużym rozmiarze $|C|$ możliwe jest rozpatrywanie wyłącznie jednoelementowych zbiorów atrybutów.

Znaczenie parametru λ jest silnie związane z obranym podejściem. W przypadku zbiorów λ, M -rankingowo znaczących wartość λ można dobrać po rozważeniu rozkładu wartości dla zbiorów rodziny $K_M(C)$. Dla zbiorów λ, M -sumacyjnie znaczących należałoby obliczyć i przeanalizować rozkład wartości funkcji val_2' dla przestrzeni 2^C . W pracy nie będziemy zajmować się tym zagadnieniem w ogólności. W poprzednim rozdziale zaprezentowaliśmy zwięzły i efektywny algorytm wyznaczania zbioru λ, l -sumacyjnie znaczącego. Zauważmy, że w tym przypadku interesujące są wyłącznie takie poziomy λ , które prowadzą do znalezienia zbioru znaczącego różnego od C .

5. Dyskretyzacja

5.1. Wprowadzenie

Dane dotyczące poziomów ekspresji mają charakter ciągły. Wiele algorytmów służących analizie danych, w tym propozycje znajdujące się w dalszej części pracy, potrzebuje wielkości o charakterze dyskretnym. Z tego względu celowe jest znalezienie dla tablicy decyzyjnej reprezentującej eksperyment tablicy o wartościach dyskretnych, która możliwie dobrze opisuje informacje zawarte w tablicy pierwotnej, a co za tym idzie również sam eksperyment. Podejście takie jest szeroko stosowane w metodach odkrywania wiedzy. W przypadku rozpatrywanej dziedziny najczęściej stosuje się dyskretyzację binarną. Przy czym dla każdego atrybutu wartość 0 oznacza ogólnie niski poziom ekspresji odpowiadającego mu genu, a wartość 1 – poziom wysoki.

5.2. Definicje

Poniższe definicje odnoszą się do dyskretyzacji globalnej. Każdy atrybut jest rozpatrywany niezależnie od pozostałych. Ignorowany jest ewentualnie niejednorodny charakter uniwersum ze względu na dany atrybut.

Niech:

- $a: U \rightarrow V_a$ będzie atrybutem ciągłym
- I_a będzie podziałem V_a na przedziały
- $\leq_{I_a} \subset I_a^2$ będzie relacją porządku w zbiorze I_a

Definicja 5.2-1

Dyskretyzacja atrybutu ciągłego a oznacza znalezienie atrybutu porządkowego $a': U \rightarrow V_a'$, takiego że:

- $a' = a \circ g \circ h$
- $g: V_a \rightarrow I_a$, taka że $\forall (a \in V_a, I \in I_a) (I = g(a) \Leftrightarrow a \in I)$
- $h: I_a \rightarrow V_a'$, taka że $\forall (I_1, I_2 \in I_a) (I_1 \leq_{I_a} I_2 \Rightarrow h(I_1) \leq h(I_2))$

Definicja 5.2-2

Dyskretyzacja zbioru atrybutów A oznacza znalezienie takiego zbioru A' , który zawiera dokładnie:

- każdy atrybut dyskretny $a \in A$
- atrybuty dyskretne uzyskane po dyskretyzacji każdego atrybutu ciągłego $a \in A$

Definicja 5.2-3

Dyskretyzacja systemu informacyjnego (U, A) oznacza znalezienie systemu (U, A') , takiego że zbiór A' jest wynikiem dyskretyzacji zbioru atrybutów A

Definicja 5.2-4

Dyskretyzacja tablicy decyzyjnej (U, A, C, D) oznacza wyznaczenie tablicy (U, A', C', D') , gdzie:

- zbiór C' jest wynikiem dyskretyzacji zbioru atrybutów C
- zbiór D' jest wynikiem dyskretyzacji zbioru atrybutów D
- $A' = C' \cup D'$

Przykład

Rozpatrzmy tablicę decyzyjną DT_2 . W celu podania przykładowej dyskretyzacji rozpatrzmy zbiór atrybutów warunkowych $\{G_1, G_2, G_3, G_4, G_5\}$ tablicy decyzyjnej. Rezultat dyskretyzacji zależy od doboru dla każdego atrybutu określonego podziału jego zbioru wartości oraz funkcji g i h .

Atrybut pierwotny	Podział przeciwdziedziny atrybutu pierwotnego I_a	Atrybut zdyskretyzowany	Przeciwdziedzina atrybutu zdyskretyzowanego V_a'
G_1	$\{<1,01;3,25>,(3,25;9,65>\}$	G_1'	$\{0,1\}$
G_2	$\{<0,03;0,53>,<0,53;1,12>,<1,12;1,53>\}$	G_2'	$\{0,1,2\}$
G_3	$\{<0,36;3,67>,<3,67;5,15>,(5,15;6,33>\}$	G_3'	$\{0,1,2\}$
G_4	$\{<1,17;5,27>,<5,27;8,34>\}$	G_4'	$\{0,1\}$
G_5	$\{<2,11;4,77>,(4,77;8,02>\}$	G_5'	$\{0,1\}$

Tab. 5.2-1. Wielkości wyznaczone przy dyskretyzacji

Zaproponowane podziały i funkcje wyznaczają jednoznacznie zbiór atrybutów $\{G_1', G_2', G_3', G_4', G_5'\}$. W ten sposób zdyskretyzowana tablica decyzyjna dla tablicy DT_2 przyjmuje postać:

$(\{H_1, H_2, H_3, H_4, H_5, H_6, H_7, H_8\}, \{G_1', G_2', G_3', G_4', G_5', d\}, \{G_1', G_2', G_3', G_4', G_5'\}, \{d\})$.

Została ona zilustrowana w tabeli 5.2-2.

	G_1'	G_2'	G_3'	G_4'	G_5'	d
H_1	0	0	0	1	1	Z
H_2	0	0	0	1	1	Z
H_3	0	0	2	0	0	Z
H_4	1	0	2	1	1	Z
H_5	1	1	0	0	0	R
H_6	1	2	0	0	0	R
H_7	1	1	0	0	0	R
H_8	1	1	0	0	0	R

Tab. 5.2-2. Tablica decyzyjna DT_2 po dyskretyzacji zgodnie z wielkościami z tabeli 4.2.2-2

5.3. Podstawowe metody

Dyskretyzacja systemu informacyjnego (U, A) sprowadza się do dyskretyzacji atrybutów ciągłych należących do zbioru A . W przypadku tablicy decyzyjnej dyskretyzacja odnosi się wyłącznie do atrybutów warunkowych.

Dostępna literatura dotycząca dziedziny nie wspomina o stosowaniu zaawansowanych metod dyskretyzacji lokalnej, dlatego dalej rozważam przypadek globalny.

Dyskretyzacja globalna atrybutu polega na zastąpieniu go atrybutem porządkowym, którego wartości odpowiadają przedziałom wartości atrybutu oryginalnego. Przyporządkowanie zachowuje relację porządku pomiędzy przedziałami. Metody sprowadzają się do wyznaczenia podziału zbioru wartości I_a atrybutu oryginalnego. Podział ten jest jednoznacznie

reprezentowany przez zbiór końców przedziałów należących do I_a , które często określa się mianem wartości progowych.

Jeśli rozpatrujemy tablicę decyzyjną, to dyskretyzacja może przebiegać bez nadzoru lub z nadzorem. Nadzór oznacza w tym wypadku uwzględnienie przy przekształceniu atrybutu jego związku z wyróżnionym atrybutem decyzyjnym. Dyskretyzacja dla każdego atrybutu wprowadza pewien podział jego zbioru wartości. Prace [23,24] wskazują, że podziały takie mogą być bezpośrednio interpretowane, jako nośniki informacji na temat klasy decyzyjnej. Podsumowując, celem metod z nadzorem jest wyznaczenie takiego nowego zestawu atrybutów, który zapewni jak najlepszą reprezentację pojęcia decyzyjnego.

Celem niniejszej pracy nie jest analiza różnych metod dyskretyzacji, dlatego ograniczam się do przedstawienia metod, które są powszechnie stosowane w badaniu ekspresji genów.

5.3.1. Równa szerokość przedziałów

Metoda jest jedną z najprostszych propozycji [10,20], pomimo to bywa często stosowana w rozważanej dziedzinie [28]. Parametrem wejściowym rozważany atrybut i liczba przedziałów w szukanym podziale I_a . Progi są ustalane arbitralnie tak, aby przedziały miały tę samą szerokość. Dużą wadą rozwiązania jest ignorowanie rozkładu wartości w przeciwdziedzinie atrybutu V_a .

Przykład

Rozważmy dyskretyzację binarną tablicy decyzyjnej DT_2 , każdy atrybut dyskretyzowany jest przy założeniu podziału jego przeciwdziedziny na dwa przedziały o jednakowej szerokości. Podziały oraz przeciwdziedziny odpowiednich atrybutów zdyskretyzowanych przedstawione zostały w tabeli 5.3.1-1, a zdyskretyzowana tablica decyzyjna w tabeli 5.3.1-2.

Atrybut pierwotny	Podział przeciwdziedziny atrybutu pierwotnego I_a	Atrybut zdyskretyzowany	Przeciwdziedzina atrybutu zdyskretyzowanego V_a'
G_1	$\{<1,01;5,33>, <5,33;9,65>\}$	G_1'	$\{0,1\}$
G_2	$\{<0,03;0,78>, <0,78;1,53>\}$	G_2'	$\{0,1\}$
G_3	$\{<0,36;3,345>, <3,345;6,33>\}$	G_3'	$\{0,1\}$
G_4	$\{<1,17;4,755>, <4,755;8,34>\}$	G_4'	$\{0,1\}$
G_5	$\{<2,11;5,065>, <5,065;8,02>\}$	G_5'	$\{0,1\}$

Tab. 5.3.1-1. Wielkości wyznaczone przy dyskretyzacji dla równej szerokości przedziałów

	G_1'	G_2'	G_3'	G_4'	G_5'	d
H_1	0	0	0	1	1	Z
H_2	0	0	0	1	1	Z
H_3	0	0	1	0	0	Z
H_4	0	0	1	1	1	Z
H_5	0	1	0	0	0	R
H_6	0	1	0	0	0	R
H_7	1	1	0	1	0	R
H_8	1	1	0	0	0	R

Tab. 5.3.1-2. Tablica decyzyjna po dyskretyzacji zgodnie z wielkościami z tab. 5.3.1-1

Otrzymana tablica zostanie wykorzystana do dalszej analizy i oznaczona przez DT_3 . Dla przejrzystości nowe atrybuty będą w następnych rozdziałach oznaczane symbolami odpowiadających im atrybutów pierwotnych.

5.3.2. Równa głębokość przedziałów

Metoda jest pewnym udoskonaleniem podziału zbioru wartości atrybutu na przedziały równej szerokości. Jest ona opisana między innymi w pracach [10,20]. Podobnie jak w poprzednim podejściu, parametrem wejściowym jest rozważany atrybut i liczba przedziałów k w szukanym podziale I_a . Dla danego systemu informacyjnego podział jest dokonywany tak, aby do każdego z przedziałów należała możliwie taka sama liczba wartości spośród zbioru wartości atrybutu. W tym przypadku progi są związane z rozkładem wartości atrybutu w jego przeciwdziedzinie.

Metoda zmierza do osiągnięcia podziału I_a zawierającego k przedziałów o długości $\frac{|a(U)|}{k}$.

Ze względu na charakter zbioru wartości atrybutu w rozpatrywanym systemie informacyjnym, może być niemożliwe osiągnięcie równej liczby wartości w każdym przedziale lub założonej liczby przedziałów.

Przykład

Rozważmy dyskretyzację binarną tablicy decyzyjnej DT_2 , każdy atrybut dyskretyzowany jest przy założeniu podziału jego przeciwdziedziny na dwa przedziały o jednakowej głębokości. Podziały oraz przeciwdziedziny odpowiednich atrybutów zdyskretyzowanych przedstawione zostały w tabeli 5.3.2-1, a zdyskretyzowana tablica decyzyjna w tabeli 5.3.2-2.

Atrybut pierwotny	Podział przeciwdziedziny atrybutu pierwotnego I_a	Atrybut zdyskretyzowany	Przeciwdziedzina atrybutu zdyskretyzowanego V_a'
G_1	$\{<1,01;1,74>, <1,74;8,17>, <8,17;9,65>\}$	G_1'	$\{0,1,2\}$
G_2	$\{<0,03;0,21>, <0,21;1,00>, <1,00;1,53>\}$	G_2'	$\{0,1,2\}$
G_3	$\{<0,36;0,91>, <0,91;5,22>, <5,22;6,33>\}$	G_3'	$\{0,1,2\}$
G_4	$\{<1,17;3,32>, <3,32;6,02>, <6,02;8,34>\}$	G_4'	$\{0,1,2\}$
G_5	$\{<2,11;3,33>, <3,33;6,44>, <6,44;8,02>\}$	G_5'	$\{0,1,2\}$

Tab. 5.3.2-1. Wielkości wyznaczone przy dyskretyzacji dla równej głębokości przedziałów

	G_1'	G_2'	G_3'	G_4'	G_5'	d
H_1	0	1	1	1	2	Z
H_2	1	0	1	2	1	Z
H_3	0	0	2	1	1	Z
H_4	0	0	2	2	2	Z
H_5	1	2	0	0	0	R
H_6	1	2	0	0	0	R
H_7	2	1	1	1	0	R
H_8	2	1	0	0	1	R

Tab. 5.3.1-2. Tablica decyzyjna DT_2 po dyskretyzacji zgodnie z wielkościami z tab. 5.3.1-1

5.3.3. Podejście zstępujące

Metoda jest jedynym prezentowanym tutaj algorytmem dyskretyzacji z nadzorem i została szczegółowo opisana w pracach [10,16]. Obok rozważanego atrybutu warunkowego analizowane są również wartości atrybutu decyzyjnego dla odpowiednich obiektów. Dodatkowe parametry wejściowe zależą od wybranego warunku stopu. Wyjściem jest zbiór wartości progowych.

Algorytm dzieli rozpatrywany zbiór obiektów na dwa podzbiory, przy czym wybór progu podziału θ zależy od zarówno od wartości atrybutu, jak i klasy obiektów. Następnie procedura wykonywana jest rekurencyjnie dla każdego z uzyskanych wzorów. Zwracana jest suma zbiorów progów z obu rekurencyjnych wywołań procedury i jednoelementowego zbioru zawierającego próg θ .

Algorytm nazywany jest zstępującym, gdyż jego działanie opiera się o rekurencyjne sformułowanie rozważanego zadania. Problem wyznaczenia podziału dla określonego zbioru wartości atrybutu można sprowadzić do problemu wyznaczenia dwuelementowego podziału tego zbioru, a następnie znalezienia rozwiązania dla każdego z dwóch zbiorów tego podziału.

Rozważmy tablicę decyzyjną $(U, A, C, \{d\})$. Schemat algorytmu wyznaczania podziału dla zbioru wartości atrybutu warunkowego $a \in C$ dla obiektów zbioru U ma postać:

```

funkcja discretize-top-down( $P, a$ )
jeśli warunek-stopu ( $P, a$ ) to
    zwróć  $\emptyset$ 
 $\theta \leftarrow$  wybierz-próg ( $P, a$ )
 $\Theta_0 \leftarrow$  dyskretyzuj-zstępująco ( $P_{a \leq \theta}, a$ )
 $\Theta_1 \leftarrow$  dyskretyzuj-zstępująco ( $P_{a > \theta}, a$ )
zwróć  $\Theta_0 \cup \{\theta\} \cup \Theta_1$ 

```

Przy czym:

- θ – próg wybrany dla zbioru P
- Θ_0, Θ_1 – zbiór progów podziału
- $P_{T(a, \theta)}$ – podzbiór obiektów P , dla których prawdziwa jest funkcja zdaniowa $T(a, \theta)$
- $T(a, \theta) = \{ a \leq \theta, a > \theta \}$

Podział wynikowy silnie zależy od warunku stopu i sposobu wybierania progu. Często obie te funkcje odwołują się do tych samych wielkości określających zbiór P , stąd poprzedzane są wstępnym etapem wyznaczania tych ich wartości.

Dla danego zbioru P możliwych jest $(|P|-1)$ różnych podziałów. Spośród nich wybierany jest ten, który dostarcza najwięcej informacji na temat pojęcia decyzyjnego. Realizująca wybór funkcja wybierz-próg zwraca wartość jednoznacznie implikującą podział. Powszechnie przyjmuje się, że dla danego podziału dobrą miarą wnoszonej niego informacji jest ważona entropia zbioru obiektów P ze względu na ten podział. Entropia względem podziału $E_{a,\theta}$ jest wyrażona poprzez dwie entropie zbiorów obiektów $P_{a\leq\theta}$ i $P_{a>\theta}$ o wartościach odpowiednio $E_{a\leq\theta}$ i $E_{a>\theta}$. Zgodnie z powyższym kryterium wybierany jest podział o minimalnej entropii.

$$E_{a,\theta}(P) = \frac{|P_{a\leq\theta}|}{|P|} E_{a\leq\theta} + \frac{|P_{a>\theta}|}{|P|} E_{a>\theta} \quad E_{T(a,\theta)}(P) = \sum_{d \in V_d} - \frac{|P_{T(a,\theta)}^d|}{|P_{T(a,\theta)}|} \log_2 \frac{|P_{T(a,\theta)}^d|}{|P_{T(a,\theta)}|}$$

Przy czym: $P_{T(a,\theta)}^d = \{P : T(a,\theta) \wedge d(P) = d\}$.

Warunek stopu na danym poziomie zagłębienia algorytmu jest silnie związany z entropiami możliwych podziałów dwuelementowych. Decyzje o podzieleniu rozważanego zbioru obiektów można podjąć posługując się pojęciem zysku informacji wynikającym z wprowadzenia podziału. Zysk $g_{a,\theta}$ jest wyznaczany na podstawie entropii zbioru niepodzielnego I i entropii względem najkorzystniejszego podziału kandydującego $E_{T(a,\theta)}$.

$$g_{a,\theta}(P) = I(P) - E_{T(a,\theta)}(P) \quad I(P) = \sum_{d \in V_d} - \frac{|U^d|}{|U|} \log_2 \frac{|U^d|}{|U|}$$

Sformułowanie warunku sprowadza się do określenia progu, od którego zysk jest uznawany za wystarczający dla rezygnacji z przeprowadzenia podziału. Najprostszym kryterium jest $g_{a,\theta} \leq 0$ lub zaproponowanej innej maksymalnej wartości, np. opierając się o zadane parametry wejściowe.

W pracy [16] zaproponowany jest algorytm zstępujący, który formułuje warunek stopu w oparciu o zasadę minimalnej długości kodu [10].

$$g_{a,\theta}(P) \leq \frac{\log_2(N-1)}{N} + \frac{\delta_{a,\theta}}{N} \quad \delta_{a,\theta}(P) = \log_2(3^k - 2) - (kI - k_{a\leq\theta}E_{a\leq\theta}(P) - k_{a>\theta}E_{a>\theta}(P))$$

Przy czym:

- $N=|P|$
- $k=|d(P)|$
- $k_{T(a,\theta)}=|d(P_{T(a,\theta)})|$

Warto dodać, że prezentowana metoda może również wskazywać na istotność atrybutów dla badanej hipotezy. Jeśli dany atrybut został zdyskretyzowany do jednego przedziału, to oznacza to, że żaden podział jego zbioru wartości nie dawał wystarczająco wysokiego wzrostu informacji na temat klasy decyzyjnej. Atrybuty takie można pominąć w dalszej analizie. W niektórych przypadkach metoda ta jest stosowana bezpośrednio, bez wstępnego zawężania zbioru atrybutów warunkowych.

Metoda zstępująca odwołująca się do pojęcia entropii i zasady minimalnej długości kodu zaproponowana w [16] została wykorzystana w pracach [27,30,31] dotyczących wyznaczania wzorców wyłaniających.

Przykład

Rozważmy dyskretyzację tablicy decyzyjnej DT_2 . Podziały oraz przeciwdziedziny odpowiednich atrybutów zdyskretyzowanych przedstawione zostały w tabeli 5.3.3-1, a zdyskretyzowana tablica decyzyjna w tabeli 5.3.3-2. Zauważmy, że poza zmianą charakteru atrybutów metoda wskazała, że atrybut G_4 jest słabo związany z hipotezą.

Atrybut pierwotny	Podział przeciwdziedziny atrybutu pierwotnego I_a	Atrybut zdyskretyzowany	Przeciwdziedzina atrybutu zdyskretyzowanego V_a'
G_1	$\{<1,01;2,455>, <2,455;9,65>\}$	G_1'	$\{0,1\}$
G_2	$\{<0,03;0,55>, <0,55;1,53>\}$	G_2'	$\{0,1\}$
G_3	$\{<0,36;1,66>, <1,66;6,33>\}$	G_3'	$\{0,1\}$
G_4	$\{<1,17;8,34>\}$	G_4'	$\{0\}$
G_5	$\{<2,11;3,725>, <3,725;8,02>\}$	G_5'	$\{0,1\}$

Tab. 5.3.3-1. Wielkości wyznaczone przy dyskretyzacji dla równej głębokości przedziałów

	G_1'	G_2'	G_3'	G_4'	G_5'	d
H_1	0	0	1	0	1	Z
H_2	0	0	1	0	1	Z
H_3	0	0	1	0	1	Z
H_4	0	0	1	0	1	Z
H_5	1	1	0	0	0	R
H_6	1	1	0	0	0	R
H_7	1	1	0	0	0	R
H_8	1	1	0	0	0	R

Tab. 5.3.3-2. Tablica decyzyjna DT_2 po dyskretyzacji zgodnie z wielkościami z tab. 5.3.3-1

6. Wyznaczanie rozróżniających grup genów

6.1. Wprowadzenie

Informację niesioną przez dany zbiór atrybutów często rozumie się przez ich zdolność do rozróżniania obiektów zgodnie z rozróżnieniem wprowadzanym przez klasę decyzyjną. Takie podejście prezentowane jest między innymi w zaproponowanej przez Zdzisława Pawłaka teorii zbiorów przybliżonych [39] oraz w publikacjach związanych z syntezą układów logicznych [23,24]. W pracy stosuję formalizm zbiorów przybliżonych, przy czym skupiam się na koncepcie reduktu.

Redukt jest minimalnym zbiorem atrybutów niezbędnych do rozróżnienia wszystkich obiektów danej tablicy decyzyjnej pomiędzy klasy decyzyjne. Podejście oparte o teorię zbiorów przybliżonych jest powszechnie wykorzystywane do redukcji systemów informacyjnych. W badaniu ekspresji genów wykorzystywane jest najczęściej do konstrukcji klasyfikatorów opartych o zbiory reguł [21].

W przypadku badania ekspresji genów konieczna jest eliminacja atrybutów nieistotnych oraz dyskretyzacja atrybutów ciągłych. Bezpośrednia analiza tablicy decyzyjnej o wysokiej wymiarowości i małym zbiorze obiektów może doprowadzić do znalezienia bardzo dużej liczby mało licznych reduktów.

Częstym zarzutem w stosunku do koncepcji reduktu jest fakt, że redukt wyznaczony dla określonej tablicy decyzyjnej w ogólności nie musi być reduktem dla tablicy o uniwersum rozszerzonym o nowe obiekty. Odpowiedzią na tę niedogodność jest propozycja reduktu dynamicznego. Wyznaczenie tego pojęcia wymaga rozważenia pewnej rodziny tablic decyzyjnych o zbiorach obiektów zawartych w uniwersum tablicy decyzyjnej. Dla każdej z tablic wyznaczane są redukty względne. Szukanymi reduktami dynamicznymi są redukty względne dla największej liczby tablic rozpatrywanej rodziny. W pracy nie zajmujemy się pojęciem reduktu dynamicznego

6.2. Elementy teorii zbiorów przybliżonych

Poniższe definicje i twierdzenia będą sformułowane w stosunku do tablic decyzyjnych, a nie systemów informacyjnych. Przy nazwach pojęć implicite zakłada się, że są one rozumiane względem decyzji. Również definicje podstawowych problemów są naturalnym przeniesieniem na przypadek tablic decyzyjnych problemów sformułowanych pierwotnie dla systemów informacyjnych.

Podstawowym źródłem dla niniejszego wstępu teoretycznego są prace [7,12]. Twierdzenia formułowane na potrzeby pracy zostaną poparte dowodami.

Definicja 6.2-1

Niech $DT=(U,A,C,\{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $B \subset C$.

Relacja nierozróżnialności $IND_{DT}(B)$ dla tablicy decyzyjnej $(U,A,C,\{d\})$ oznacza dwuargumentową relację w zbiorze U generowaną przez zbiór atrybutów B , która ma postać $IND_{DT}(B) = \{(u,v) \in U \times U : (\forall a \in B : a(u) = a(v)) \vee d(u) = d(v)\}$.

Definicja 6.2-2

Macierz rozróżnialności dla tablicy decyzyjnej $(U, A, C, \{d\})$ oznacza kwadratową macierz M o rozmiarze $n \times n$, taką że: $M_{ij} = \{a \in C : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\}$.

Podsumowując, każdy element macierzy jest zbiorem tych atrybutów, które mają różną wartość dla wybranych obiektów z uniwersum. W szczególności na przekątnej macierzy znajdują się zbiory puste. Ze względu na naturalną symetryczność relacji $\neq$ w przeciwdziedzinach atrybutów powyższe struktury zawierają redundancję informacji na temat rozróżnialności obiektów. Macierz jest symetryczna względem swojej przekątnej, zamienienie indeksów i oraz j daje w wyniku identyczne struktury.

Przykład

Dla przybliżenia pojęć będziemy w tym rozdziale rozważać tablicę decyzyjną DT_3 otrzymaną w rozdziale 5.3.1.

	G ₁	G ₂	G ₃	G ₄	G ₅	D
H ₁	0	0	0	1	1	Z
H ₂	0	0	0	1	1	Z
H ₃	0	0	1	0	0	Z
H ₄	0	0	1	1	1	Z
H ₅	0	1	0	0	0	R
H ₆	0	1	0	0	0	R
H ₇	1	1	0	1	0	R
H ₈	1	1	0	0	0	R

Tab. 6.2-1. Tablica decyzyjna DT_3

Przykładowe relacje nierozróżnialności dla tablicy DT_3 dla różnych zbiorów atrybutów:

- $IND_{DT}(C) = \{(H_1, H_1), (H_1, H_2), (H_1, H_3), (H_1, H_4), (H_2, H_1), (H_2, H_2), (H_2, H_3), (H_2, H_4), (H_3, H_1), (H_3, H_2), (H_3, H_3), (H_3, H_4), (H_4, H_1), (H_4, H_2), (H_4, H_3), (H_4, H_4), (H_5, H_5), (H_5, H_6), (H_5, H_7), (H_5, H_8), (H_6, H_5), (H_6, H_6), (H_6, H_7), (H_6, H_8), (H_7, H_5), (H_7, H_6), (H_7, H_7), (H_7, H_8), (H_8, H_5), (H_8, H_6), (H_8, H_7), (H_8, H_8)\}$
- $IND_{DT}(\{G_5\}) = IND_{DT}(C) \cup \{(H_3, H_5), (H_3, H_6), (H_3, H_7), (H_3, H_8), (H_5, H_3), (H_6, H_3), (H_7, H_3), (H_8, H_3)\}$

Macierz rozróżnialności dla tablicy DT_3 jest przedstawiona w tabeli 6.2-2. Dla przejrzystości w komórkach tabeli nie zostały umieszczone symbole atrybutów, lecz ich indeksy. Przykładowo zbiór $\{2, 4, 5\}$ opisuje zbiór atrybutów $\{G_2, G_4, G_5\}$.

	H ₁	H ₂	H ₃	H ₄	H ₅	H ₆	H ₇	H ₈
H ₁	∅	∅	∅	∅	{2,4,5}	{2,4,5}	{1,2,5}	{1,2,4,5}
H ₂	∅	∅	∅	∅	{2,4,5}	{2,4,5}	{1,2,5}	{1,2,4,5}
H ₃	∅	∅	∅	∅	{2,3}	{2,3}	{1,2,3,4}	{1,2,3}
H ₄	∅	∅	∅	∅	{2,3,4,5}	{2,3,4,5}	{1,2,3,5}	{1,2,3,4,5}
H ₅	{2,4,5}	{2,4,5}	{2,3}	{2,3,4,5}	∅	∅	∅	∅
H ₆	{2,4,5}	{2,4,5}	{2,3}	{2,3,4,5}	∅	∅	∅	∅
H ₇	{1,2,5}	{1,2,5}	{1,2,3,4}	{1,2,3,5}	∅	∅	∅	∅
H ₈	{1,2,4,5}	{1,2,4,5}	{1,2,3}	{1,2,3,4,5}	∅	∅	∅	∅

Tab. 6.2-2. Macierz rozróżnialności dla tablicy DT_3

Łatwo można dostrzec, że relacja nierozróżnialności zawiera wszystkie te pary obiektów (i,j) , dla których $M_{ij} = \emptyset$.

Definicja 6.2-3

Niech $(U,A,C,\{d\})$ będzie tablicą decyzyjną, M jej macierzą rozróżnialności i zbiór atrybutów $P \subset C$.

Zbiór atrybutów P **pokrywa** parę obiektów (u_i, u_j) macierzy M , jeśli $P \cap M_{i,j} \neq \emptyset$.

Definicja 6.2-4

Niech $(U,A,C,\{d\})$ będzie tablicą decyzyjną, M jej macierzą rozróżnialności i zbiór atrybutów $P \subset C$.

Zbiór pokrywający dla zbioru atrybutów P oznacza zbiór $\{(u_i, u_j) \in U \times U : P \cap M_{i,j} \neq \emptyset\}$.

Zbiór pokrywający oznaczamy przez $CS(P)$.

Przykład

Analizując macierz rozróżnialności dla DT_3 możemy powiedzieć, że:

- Zbiór atrybutów $\{G_2, G_3\}$ pokrywa między innymi pary obiektów $(H_3, H_5), (H_8, H_4)$
- Zbiór atrybutów $\{G_2, G_4, G_5\}$ pokrywa między innymi pary obiektów $(H_2, H_8), (H_8, H_4)$

Przejrzenie całej macierzy pozwala na wyznaczenie zbiorów pokrywających, przykładowo:

- $CS(\{G_1\}) = \{(H_1, H_7), (H_1, H_8), (H_2, H_7), (H_2, H_8), (H_3, H_7), (H_3, H_8), (H_4, H_7), (H_4, H_8), (H_7, H_1), (H_8, H_1), (H_7, H_2), (H_8, H_2), (H_7, H_3), (H_8, H_3), (H_7, H_4), (H_8, H_4)\}$
- $CS(\{G_2, G_3\}) = U \times U$

Definicja 6.2-5

Niech $DT = (U,A,C,\{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $P \subset C$.

Atrybut $a \in C$ jest:

- **zbędny** w zbiorze P , jeśli $IND_{DT}(C) = IND_{DT}(C \setminus \{a\})$
- **niezbędny** w zbiorze P , jeśli nie jest zbędny

Definicja 6.2-6

Niech $DT = (U,A,C,\{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $P \subset C$.

Zbiór atrybutów jest **niezależny** w DT , jeśli każdy atrybut $a \in P$ jest niezbędny.

Przykład

Atrybut G_4 jest zbędny dla zbioru atrybutów C , gdyż po jego usunięciu zbiór ten wciąż tak samo rozróżnia obiekty pomiędzy klasy decyzyjne, $IND_{DT_3}(C) = IND_{DT_3}(C \setminus \{G_4\})$. Dla zbioru atrybutów $\{G_1, G_2\}$ atrybut G_1 jest zbędny, a G_2 – niezbędny.

Fakty te wynikają z wartości następujących relacji rozróżnialności:

$$\begin{aligned} IND_{DT_3}(C) &= \{(H_1, H_1), (H_1, H_2), (H_1, H_3), (H_1, H_4), (H_2, H_1), (H_2, H_2), (H_2, H_3), (H_2, H_4), \\ &\quad (H_3, H_1), (H_3, H_2), (H_3, H_3), (H_3, H_4), (H_4, H_1), (H_4, H_2), (H_4, H_3), (H_4, H_4), \\ &\quad (H_5, H_5), (H_5, H_6), (H_5, H_7), (H_5, H_8), (H_6, H_5), (H_6, H_6), (H_6, H_7), (H_6, H_8), \\ &\quad (H_7, H_5), (H_7, H_6), (H_7, H_7), (H_7, H_8), (H_8, H_5), (H_8, H_6), (H_8, H_7), (H_8, H_8)\} \\ IND_{DT_3}(\{G_1, G_2, G_3, G_4\}) &= IND_{DT_3}(C) \\ IND_{DT_3}(\{G_1, G_2\}) &= IND_{DT_3}(C) \\ IND_{DT_3}(\{G_1\}) &= IND_{DT_3}(C) \cup \{(H_1, H_5), (H_1, H_6), (H_2, H_5), (H_2, H_6), \\ &\quad (H_3, H_5), (H_3, H_6), (H_4, H_5), (H_4, H_6), (H_5, H_1), (H_5, H_2), (H_5, H_3), (H_5, H_4), \\ &\quad (H_6, H_1), (H_6, H_2), (H_6, H_3), (H_6, H_4)\} \\ IND_{DT_3}(\{G_2\}) &= IND_{DT_3}(C) \end{aligned}$$

Definicja 6.2-7

Niech $DT = (U, A, C, \{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $P \subset C$.

Zbiór rozróżniający zbioru atrybutów P w tablicy DT jest równy Q , jeśli:

- $Q \subset P$
- $IND_{DT}(P) = IND_{DT}(Q)$

Definicja 6.2-8

Niech $DT = (U, A, C, \{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $P \subset C$.

Nadredukt zbioru atrybutów P w tablicy DT jest równy Q , jeśli:

- Q jest zbiorem rozróżniającym zbioru atrybutów P
- zbiór atrybutów Q zawiera atrybuty zbędne

Definicja 6.2-9

Niech $DT = (U, A, C, \{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $P \subset C$.

Redukt zbioru atrybutów P w tablicy DT jest równy Q , jeśli:

- Q jest zbiorem rozróżniającym zbioru atrybutów P
- zbiór atrybutów Q jest niezależny w DT

Zbiór wszystkich reduktów zbioru atrybutów P w tablicy decyzyjnej DT oznaczamy przez $RED_{DT}(P)$.

Przykład

Zbiorami rozróżniającymi zbioru atrybutów C dla tablicy DT_3 są między innymi:

$\{G_1, G_2\}, \{G_1, G_3, G_4, G_5\}, \{G_2\}, \{G_3, G_5\}, \{G_1, G_3, G_4\}$.

Przy czym:

- $\{G_1, G_2\}$ jest nadreduktem, gdyż atrybut G_1 jest zbędny
- $\{G_1, G_3, G_4, G_5\}$ jest nadreduktem, gdyż atrybut G_5 jest zbędny
- $RED_{DT}(C) = \{\{G_2\}, \{G_3, G_5\}, \{G_1, G_3, G_4\}\}$

Definicja 6.2-10

Niech $DT=(U,A,C,\{d\})$ będzie tablicą decyzyjną i zbiór atrybutów $P \subset C$.

Rdzeń zbioru atrybutów P w tablicy DT , jest zbiorem Q wszystkich niezbędnych atrybutów w zbiorze P , przy czym $Q \subset P$.

Rdzeń zbioru atrybutów P w tablicy decyzyjnej DT oznaczamy przez $CORE_{DT}(P)$.

Twierdzenie 6.2-1 [7]

Jeśli $DT=(U,A,C,\{d\})$ jest tablicą decyzyjną i $P \subset C$, to $CORE_{DT}(P) = \bigcap_{R \in RED_{DT}(P)} R$.

Przykład

Rdzeń zbiorów atrybutów dla tablicy DT_3 :

- $CORE_{DT}(C) = \emptyset$
- $CORE_{DT}(\{G_1, G_3, G_4, G_5\}) = \{G_3\}$

Twierdzenie 6.2-2 (oszacowanie liczby reduktów, [7])

Jeśli $DT=(U,A,C,\{d\})$ jest tablicą decyzyjną i $m=|C|$, to $1 \leq card(RED_{DT}(C)) \leq \left\lceil \left\lfloor \frac{m}{2} \right\rfloor \right\rceil$.

Przykład

Twierdzenie jest prawdziwe dla zbioru reduktów tablicy DT_3 :

$$1 \leq card(RED_{DT}(C)) = 3 \leq \left\lceil \left\lfloor \frac{5}{2} \right\rfloor \right\rceil = \left\lceil 2 \right\rceil = 3$$

Definicja 6.2-11

Redukcja tablicy decyzyjnej (U,A,C,D) do zbioru atrybutów warunkowych $C' \subset C$ oznacza znalezienie tablicy decyzyjnej $(U, C' \cup D, C', D)$.

Definicja 6.2-12

Problem wyznaczania wszystkich reduktów (Reduct Set Problem) dla tablicy decyzyjnej $DT=(U,A,C,\{d\})$ polega na znalezieniu zbioru $RED_{DT}(C)$.

Twierdzenie 6.2-3 (złożoność problemu wyznaczania wszystkich reduktów, [7])

Problem wyznaczania wszystkich reduktów jest NP-trudny.

Definicja 6.2-13

Problem wyznaczania minimalnego reduktu (Minimal Reduct Problem) dla tablicy decyzyjnej $DT=(U,A,C,\{d\})$ polega na znalezieniu zbioru $R \subset C$, takiego że $card(R) = \min_{B \in RED_{DT}(C)} \{card(B)\}$.

Twierdzenie 6.2-4 (złożoność problemu wyznaczania minimalnego reduktu, [7])

Problem wyznaczania minimalnego reduktu jest NP-trudny.

Ze względu na istnienie wielu interesujących algorytmów proponowanych dla różnych problemów NP-trudnych warto zastanowić się nad sformułowaniem powyższych problemów w języku dziedzin, które nie powstały na gruncie teorii zbiorów przybliżonych.

Eleganckiego aparatu dla wyrażenia części wprowadzonych wyżej pojęć dostarcza algebra Boole'a. W szczególności problem wyznaczania wszystkich reduktów może zostać sprowadzony do problemu wyznaczania wszystkich implikantów pierwszych pewnej funkcji.

Definicja 6.2-14

Funkcja boolowska oznacza funkcję postaci $f : \{0,1\}^n \rightarrow \{0,1\}$.

Definicja 6.2-15

Implikant funkcji boolowskiej $f(x_1, \dots, x_n)$ oznacza term $g = x_{i_1}^{(\varphi_1)} x_{i_2}^{(\varphi_2)} \dots x_{i_r}^{(\varphi_r)}$, gdzie $\varphi_k \in \{0,1\}$, $x_{i_k}^{(0)} = \bar{x}_{i_k}$, $x_{i_k}^{(1)} = x_{i_k}$, $i_k \in \{1, 2, \dots, n\}$, $k = 1, \dots, r$, jeżeli jest prawdziwa dla każdego wartościowania V nierówność $g(V) \leq f(V)$.

Definicja 6.2-16

Implikant pierwszy oznacza implikant g , z którego nie można usunąć żadnej zmiennej.

Definicja 6.2-17

Funkcję boolowską f nazywamy **monotoniczną**, jeśli $(\alpha_1, \dots, \alpha_n) \leq (\beta_1, \dots, \beta_n)$ (to znaczy $\alpha_i \leq \beta_i$ dla każdego i) implikuje $f(\alpha_1, \dots, \alpha_n) \leq f(\beta_1, \dots, \beta_n)$ dla każdego $(\alpha_1, \dots, \alpha_n), (\beta_1, \dots, \beta_n) \in \{0,1\}^n$.

Twierdzenie 6.2-5 ([12])

Każda funkcja monotoniczna da się przedstawić w postaci alternatywy wszystkich implikantów pierwszych tej funkcji.

Zbiór implikantów pierwszych danej funkcji monotonicznej będziemy oznaczać przez PI_f .

Twierdzenie 6.2-6 (prawa pochłaniania, [12])

Dla każdych zdań logicznych a i b zachodzi:

- $a \wedge a = a$
- $a \wedge (a \vee b) = a$.

Definicja 6.2-18

Funkcja rozróżnialności f_{DT} tablicy decyzyjnej $DT = (U, \{a_1, \dots, a_m\}, \{a_1, \dots, a_n\}, \{d\})$ oznacza funkcję boolowską zmiennych $B = \{b_1, \dots, b_n\}$ odpowiadających poszczególnym atrybutom $a_1, \dots, a_n$ zdefiniowaną jako: $\bigwedge_{\substack{i,j=1,\dots,n \\ M_{ij} \neq \emptyset}} \bigvee_{b_k \in M_{ij}} b_k$, gdzie $\overline{M_{ij}} = \{b_k \in B : a_k \in M_{ij}\}$.

Funkcja przyporządkowuje wybranemu zbiorowi atrybutów zadanemu przez ustawienie na 1 odpowiednich argumentów $b_1, \dots, b_n$ wartość 1 , jeśli zbiór ten pozwala rozróżnić wszystkie obiekty U . Poniższe twierdzenie podkreśla fakt, że macierz rozróżnialności może być traktowana jako reprezentacja funkcji rozróżnialności i odwrotnie.

Przykład

Funkcja rozróżnialności dla tablicy DT_3 jest wyznaczana poprzez przejście macierzy rozróżnialności. Ponieważ macierz jest symetryczna, więc można zrezygnować z termów odpowiadających zbiorom atrybutów z M_{ij} , dla $i > j$.

$$\begin{aligned}
f_{DT_3} = & (b_2 \vee b_4 \vee b_5) \wedge (b_2 \vee b_4 \vee b_5) \wedge (b_1 \vee b_2 \vee b_5) \wedge (b_1 \vee b_2 \vee b_4 \vee b_5) \wedge \\
& (b_2 \vee b_4 \vee b_5) \wedge (b_2 \vee b_4 \vee b_5) \wedge (b_1 \vee b_2 \vee b_5) \wedge (b_1 \vee b_2 \vee b_4 \vee b_5) \wedge \\
& (b_2 \vee b_3) \wedge (b_2 \vee b_3) \wedge (b_1 \vee b_2 \vee b_3 \vee b_4) \wedge (b_1 \vee b_2 \vee b_3) \wedge \\
& (b_2 \vee b_3 \vee b_4 \vee b_5) \wedge (b_2 \vee b_3 \vee b_4 \vee b_5) \wedge (b_1 \vee b_2 \vee b_3 \vee b_5) \wedge (b_1 \vee b_2 \vee b_3 \vee b_4 \vee b_5)
\end{aligned}$$

Wyznaczona funkcja może być jeszcze uproszczona zgodnie z prawami pochłaniania.

Twierdzenie 6.2-7 ([12])

Niech $(U, A, C, \{d\})$ będzie tablicą decyzyjną, a f jej funkcją rozróżnialności. Dla dowolnego $B \subset C$ zachodzi równoważność $IND_B = IND_C \Leftrightarrow f(B) = 1$.

Twierdzenie 6.2-8 ([12])

Niech DT będzie tablicą decyzyjną, a f jej funkcją rozróżnialności. Zachodzi równoważność $(a_{i_1}, \dots, a_{i_p}) \in RED_{DT}(C) \Leftrightarrow (b_{i_1} \dots b_{i_p}) \in PI_f$.

Innego spojrzenia dostarcza teoria zbiorów. Pojęcie reduktu często odnoszone jest do pojęcia minimalnego pokrycia zbioru lub zbiorów trafieniowych. Tutaj skupimy się na przypadku zbiorów trafieniowych. Ponadto wyrazimy problem wyznaczania wszystkich reduktów tablicy decyzyjnej za pomocą równoważnego problemu wyznaczania wszystkich minimalnych zbiorów trafieniowych.

Definicja 6.2-19

Niech S będzie dowolnym zbiorem skończonym, a F dowolną rodziną podzbiorów S .

Zbiór trafieniowy dla F oznacza zbiór H , taki że $\forall D \in F : D \cap H \neq \emptyset$.

Definicja 6.2-20

Niech S będzie dowolnym zbiorem skończonym, a F dowolną rodziną podzbiorów S .

Zbiór trafieniowy H dla F jest **minimalny**, jeśli żaden podzbiór właściwy H nie jest trafieniowy.

Zbiór wszystkich minimalnych zbiorów trafieniowych dla rodziny F podzbiorów zbioru S oznaczamy przez $HS(F)$.

Przykład

Niech:

- $S = \{a, b, c, d\}$
- $F = \{\{a, b, d\}, \{b, c\}, \{a, b, c\}\}$

Zbiorami trafieniowymi są między innymi: $\{a, b, c, d\}, \{a, b, c\}, \{a, c, d\}, \{b\}$.

Rodzina minimalnych zbiorów trafieniowych wynosi: $HS(F) = \{\{b\}, \{c, d\}\}$.

Definicja 6.2-21

Problem wyznaczania wszystkich minimalnych zbiorów trafieniowych (Minimal Hitting-Set Problem) dla rodziny F podzbiorów zbioru S polega na znalezieniu zbioru $HS(F)$.

Twierdzenie 6.2-9

Niech:

- $DT=(U,A,C,\{d\})$ będzie tablicą decyzyjną
- M oznacza macierz rozróżnialności tablicy DT
- zbiór atrybutów $P \subset C$
- rodzina zbiorów atrybutów $F = \{M_{ij} : i, j = 1, \dots, m \wedge M_{ij} \neq \emptyset\}$

P jest zbiorem trafieniowym dla F , wtedy i tylko wtedy, gdy P jest zbiorem rozróżniającym w DT .

Dowód:

Na wstępie zauważmy, że zgodnie z prawami de Morgana dla każdej pary $(u_i, u_j) \in U \times U$ prawdziwa jest równoważność $(\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j) \Leftrightarrow M_{ij} = \emptyset$ (*).

Niech P będzie zbiorem trafieniowym dla rodziny F . Rodzina F zawiera zbiory atrybutów, które rozróżniają wszystkie pary obiektów różnych ze względu na klasę decyzyjną. Zgodnie z definicją 6.2-19 mamy $\forall D \in F : D \cap P \neq \emptyset$. Oznacza to, że zbiór P ma niepustą część wspólną z każdym elementem rodziny F , część ta zawiera przynajmniej jeden rozróżniający atrybut. Innymi słowy $\forall (u_i, u_j) \in U \times U (M_{ij} \neq \emptyset \Rightarrow (\exists a \in P : a(u_i) \neq a(u_j)) \wedge d(u_i) \neq d(u_j))$ (**).

$$\begin{aligned} IND_{DT}(P) &= \{(u_i, u_j) \in U \times U : (\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} = \\ &= \{(u_i, u_j) \in U \times U : M_{ij} = \emptyset \wedge (\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} \cup \\ &= \{(u_i, u_j) \in U \times U : M_{ij} \neq \emptyset \wedge (\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} = \\ &= \{(u_i, u_j) \in U \times U : ((\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)) \wedge (\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} \cup \\ &= \{(u_i, u_j) \in U \times U : M_{ij} \neq \emptyset \wedge (\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} = \\ &= \{(u_i, u_j) \in U \times U : (\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} \cup \emptyset = IND_{DT}(C) \end{aligned}$$

Drugi składnik jest pusty na mocy warunku (**). Ze względu na powyższe i na założenie, że $P \subset C$, zgodnie z definicją 6.2-7, zbiór P jest zbiorem rozróżniającym w DT .

Założmy teraz, że P jest zbiorem rozróżniającym w DT . Zgodnie z definicją 6.2-7 prawdziwe jest $IND_{DT}(P) = IND_{DT}(C)$, co zgodnie z definicją 6.2-1 można zapisać jako

$$\begin{aligned} &\{(u_i, u_j) \in U \times U : (\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} = \\ &\{(u_i, u_j) \in U \times U : (\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} \end{aligned}$$

Wnioskujemy, że dla każdej pary $(u_i, u_j) \in U \times U$ prawdziwe są równoważności:

- $(\forall a \in P : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j) \Leftrightarrow (\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)$
- $(\exists a \in P : a(u_i) \neq a(u_j)) \wedge d(u_i) \neq d(u_j) \Leftrightarrow (\exists a \in C : a(u_i) \neq a(u_j)) \wedge d(u_i) \neq d(u_j)$ (***)

Niech M_{ij} będzie dowolnym elementem rodziny F . Z założenia o postaci rodziny F mamy, że $M_{ij} \neq \emptyset$, stąd zachodzi $(\exists a \in C : a(u_i) \neq a(u_j)) \wedge d(u_i) \neq d(u_j)$, a na podstawie równoważności (***) zachodzi również $(\exists a \in P : a(u_i) \neq a(u_j)) \wedge d(u_i) \neq d(u_j)$.

Zgodnie z definicją 6.2-2 i powyższymi faktami mamy

$$M_{ij} \cap P = \{a \in C : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\} \cap P = \{a \in P : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\} \neq \emptyset.$$

Ponieważ element M_{ij} był dowolnie wybrany z rodziny F , więc na mocy definicji 6.2-19 zbiór P jest zbiorem trafieniowym w F .

□

Twierdzenie 6.2-10

Niech:

- $DT=(U,A,C,\{d\})$ będzie tablicą decyzyjną
- M oznacza macierz rozróżnialności tablicy DT
- zbiór atrybutów $P \subset C$
- rodzina zbiorów atrybutów $F = \{M_{ij} : i, j = 1, \dots, m \wedge M_{ij} \neq \emptyset\}$

Zachodzi:

- 1) P jest minimalnym zbiorem trafieniowym dla F , wtedy i tylko wtedy, gdy P jest redukt w DT
- 2) $RED_{DT}(C)=HS(F)$

Dowód:

1) Niech P będzie minimalnym zbiorem trafieniowym. Oznacza to, że jest on zbiorem trafieniowym, takim że żaden jego podzbiór właściwy nie jest zbiorem trafieniowym. Zgodnie z twierdzeniem 6.2-9 zbiór P jest zbiorem rozróżniającym w tablicy DT (*).

Niech X będzie dowolnym zbiorem atrybutów takim, że: $X \subset P \wedge X \neq P$. Zbiór X nie jest zbiorem trafieniowym, stąd istnieje takie $M_{ij} \neq \emptyset$, że $M_{ij} \cap X = \emptyset$. Zgodnie z definicją 6.2-2 ostatni warunek można zapisać, jako:

$$\{a \in C : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\} \cap X = \{a \in X : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\} = \emptyset$$

i równoważnie: $(\forall a \in X : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)$. Zgodnie z definicją 6.2-1 mamy $(u_i, u_j) \in IND_{DT}(X)$ (**).

Ponieważ $M_{ij} \neq \emptyset$ oraz zgodnie z definicjami 6.2-1 i 6.2-2 zachodzi

$$IND_{DT}(C) = \{(u_i, u_j) \in U \times U : (\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} = \{(u_i, u_j) \in U \times U : M_{ij} = \emptyset\},$$

więc $(u_i, u_j) \notin IND_{DT}(C)$ (***)

Ponieważ zbiór P jest rozróżniający w DT (*), więc $IND_{DT}(P) = IND_{DT}(C)$ (****).

Zgodnie z (**), (***) i (****) zachodzi $IND_{DT}(X) \neq IND_{DT}(P)$. Zbiór X został wybrany w sposób dowolny, więc zbiór P jest niezależny. Ponieważ zbiór P jest rozróżniający i niezależny w DT , więc na mocy definicji 6.2-9 jest on redukt w DT .

Założmy teraz, że P jest redukt w DT . Jest on zbiorem rozróżniającym i niezależnym w DT . Zgodnie z twierdzeniem 6.2-9 zbiór P jest zbiorem trafieniowym dla rodziny F .

Niech X będzie dowolnym zbiorem atrybutów takim, że: $X \subset P \wedge X \neq P$. Ze względu na niezależność zbioru P mamy, że $IND_{DT}(X) \neq IND_{DT}(P)$. Ponieważ P jest zbiorem rozróżniającym, więc $IND_{DT}(P) = IND_{DT}(C)$, a stąd $IND_{DT}(X) \neq IND_{DT}(C)$. Ponieważ $X \subset C$, więc zgodnie z definicją 6.2-1 $IND_{DT}(C) \subset IND_{DT}(X)$.

W ten sposób warunek $IND_{DT}(X) \neq IND_{DT}(C)$ oznacza, że istnieje taka para obiektów (u_i, u_j) , że $(u_i, u_j) \in IND_{DT}(X)$ i $(u_i, u_j) \notin IND_{DT}(C)$.

Ponieważ z definicji 6.2-1 i 6.2.2 mamy:

$$IND_{DT}(C) = \{(u_i, u_j) \in U \times U : (\forall a \in C : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)\} = \{(u_i, u_j) \in U \times U : M_{ij} = \emptyset\},$$

więc $M_{ij} \neq \emptyset$.

Zgodnie z definicją 6.2-1 mamy $(\forall a \in X : a(u_i) = a(u_j)) \vee d(u_i) = d(u_j)$ równoważnie możemy zapisać: $(\neg \exists a \in X : a(u_i) \neq a(u_j)) \wedge d(u_i) \neq d(u_j)$. Zgodnie z tym warunkiem mamy:

$$M_{ij} \cap X = \{a \in C : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\} \cap X = \{a \in X : a(u_i) \neq a(u_j) \wedge d(u_i) \neq d(u_j)\} = \emptyset.$$

Ponieważ $M_{ij} \neq \emptyset$, więc $M_{ij} \in F$. Zbiór X nie jest zbiorem trafieniowym, gdyż $M_{ij} \cap X = \emptyset$.

Ponieważ zbiór X został wybrany w sposób dowolny, więc zbiór H jest minimalnym zbiorem trafieniowym dla rodziny F .

2) Teza natychmiast wynika z punktu 1).

□

6.3. Realizacja wybranych pojęć

6.3.1. Struktury rozróżnialności obiektów

Algorytmy szukające reduktów w tablicy decyzyjnej zwykle wykorzystują te same struktury jednoznacznie wynikające z rozpatrywanej tablicy.

Podstawową strukturą określającą, jakie atrybuty pozwalają na rozróżnienie par obiektów jest macierz rozróżnialności.

Korzystając z symetrii przy wyliczaniu całej macierzy rozróżnialności można ograniczyć się wyłącznie do przejrzenia wszystkich dwuelementowych kombinacji obiektów z U . Dla każdej kombinacji obiektów z różnych klas jest wyznaczany zbiór atrybutów, na których te obiekty się różnią. W implementacji nadmiarowe informacje w ogóle nie są przechowywane. Bez względu na ewentualną optymalizację etap ma złożoność czasową i pamięciową $O(nm^2)$.

Schemat wyznaczania macierzy rozróżnialności

```
funkcja find-discernibility-matrix
dla każdego  $i$  od 1 do  $m$ 
     $M_{i,i} \leftarrow \emptyset$ 
dla każdego  $i$  od 1 do  $m$ 
    dla każdego  $j$  od  $i + 1$ 
        jeśli  $d(u_i) \neq d(u_j)$ 
             $M_{i,j} \leftarrow \emptyset$ 
            dla każdego  $k$  od 1 do  $n$ 
                jeśli  $a_k(u_i) \neq a_k(u_j)$  to
                     $M_{i,j} \leftarrow M_{i,j} \cup \{a_k\}$ 
             $M_{j,i} \leftarrow M_{i,j}$ 
        inaczej  $M_{i,j} \leftarrow \emptyset$ 
zwróć  $M$ 
```

6.3.2. Funkcja rozróżnialności

Dla danej tablicy decyzyjnej i jej macierzy rozróżnialności możemy skorzystać z konstruktywnego charakteru definicji 6.2-18 i wyznaczyć funkcję rozróżnialności.

Z twierdzenia 6.2-8 wynika, że każdy implikant pierwszy odpowiada reduktowi rozpatrywanej tablicy decyzyjnej. Fakt ten pozwala sprowadzić problem wyznaczania zbioru

wszystkich reduktów do problemu minimalizacji funkcji boolowskiej. Idea ta została wykorzystana w prezentowanym poniżej podejściu dokładnym.

Ponieważ funkcja rozróżnialności ma postać koniunkcyjno-alternatywną, więc zgodnie z definicją 6.2-17 jest ona monotoniczna. Na podstawie twierdzenia 6.2-5 można dla niej znaleźć równoważną funkcję zbudowaną z alternatywy implikantów pierwszych funkcji wyjściowej. Ze względów efektywnościowych etap minimalizacji funkcji warto poprzedzić redukcją funkcji poprzez usunięcie czynników:

- zawsze prawdziwych, które odpowiadają pustym elementom macierzy nierozróżnialności
- zbędnych, poprzez zastosowanie praw pochłaniania (twierdzenie 6.2-6)

Krok ten ma złożoność czasową $O(n^2m^2)$.

6.4. Metody znajdowania wszystkich reduktów

6.4.1. Podejście dokładne

Algorytm dokładny znajduje wszystkie redukty danej tablicy decyzyjnej. Jego najistotniejszym elementem jest minimalizacja funkcji rozróżnialności.

Schemat algorytmu:

funkcja exact-reduct-finder

$M \leftarrow$ znajdź macierz rozróżnialności

$f \leftarrow$ wyznacz funkcję rozróżnialności na podstawie macierzy M

$f' \leftarrow$ zredukuj postać funkcji rozróżnialności f

$d \leftarrow$ liczba czynników funkcji f'

$LR_0 = \emptyset$

dla każdego i **od** 1 **do** d

$C_i \leftarrow$ zbiór atrybutów odpowiada i -temu czynnikowi funkcji f'

$S_i = \{R \in LR_{i-1} : R \cap C_i \neq \emptyset\}$

$$T_i = \begin{cases} \{a\} \cup R : a \in C_i, R \cap C_i = \emptyset, R \in LR_{i-1} & LR_{i-1} \neq \emptyset \\ \{a\} : a \in C_i & LR_{i-1} = \emptyset \end{cases}$$

$LR_i = S_i \cup T_i$

z elementów LR_d usuń atrybuty zbędne.

z LR_d usuń elementy powtarzające się.

zwróć LR_d

Algorytm zwraca zbiór $RED_{DT}(C)$. Należy zaznaczyć, że metoda posiada istotne wady. Po pierwsze może być zastosowana wyłącznie do niedużych tablic decyzyjnych, tzn. kilkaset obiektów i kilkadziesiąt atrybutów. Dla większych zbiorów danych okazuje się nieefektywna. Po drugie otrzymanie jakichkolwiek reduktów tablicy wymaga zakończenia całego algorytmu.

Na płaszczyźnie rozwiązań dokładnych zmagania z pierwszym zarzutem przenoszą obliczenia do dziedziny algorytmów równoległych. Drugi problem może zostać rozwiązany przez pewną modyfikację zaprezentowanego wyżej podejścia, która pozwala na sukcesywne powiększanie zbioru znalezionych reduktów przez czas działania procedury.

6.4.2. Podejście zachłanne

Jedne z najprostszych metod stosowanych do wyznaczania reduktów mają charakter zachłanny.

Niech będzie dana tablica decyzyjna $(U, A, C, \{d\})$. Procedury wyznaczają pewien zbiór rozróżniający analizując kolejno atrybuty warunkowe tablicy zgodnie z ustalonym porządkiem. Kolejność przeglądania atrybutów może być modelowana przez permutację atrybutów $P = (a_{i_1}, \dots, a_{i_n})$. Permutacja ta jednoznacznie określa zbiór wynikowy.

Pierwsza metoda startuje ze zbioru pustego i dodaje do niego atrybuty zgodnie z permutacją P do momentu, gdy zbiór stanie się rozróżniający. Istotną własnością algorytmu jest fakt, że niekoniecznie cała permutacja jest wykorzystywana do znalezienia zbioru rozróżniającego.

Schemat metody:

```
funkcja greedy-adding-differentiating-set-finder  
B =  $\emptyset$   
i = 1  
dopóki CS(B)  $\neq$  CS(C)  
    x  $\leftarrow$  Pi  
    B  $\leftarrow$  B  $\cup$  {x}  
    i  $\leftarrow$  i + 1  
zwróć B
```

Druga metoda rozpoczyna działanie z zadany zbiorem początkowym i zgodnie z permutacją P odejmuje dany atrybut, jeśli jest on zbędny. W tym wypadku do redukcji jest wykorzystywana cała permutacja, a otrzymany zbiór rozróżniający jest zawsze reduktem.

Schemat metody:

```
funkcja greedy-subtracting-differentiating-set-finder  
B = C  
i = 1  
dopóki i  $\leq$  |C|  
    x  $\leftarrow$  Pi  
    jeśli CS(B \ {x}) = CS(C) to  
        B  $\leftarrow$  B \ {x}  
    i  $\leftarrow$  i + 1  
zwróć B
```

Warto zaznaczyć, że jeśli dana tablica decyzyjna posiada wiele reduktów o małym rozmiarze i nie jest konieczne otrzymanie optymalnego zbioru rozróżniającego, to metody te okazują się bardzo przydatne.

Wielokrotne powtórzenie wybranej z metod pozwala na wyznaczenie pewnego zbioru zbiorów rozróżniających. Zauważmy jednak, że bezpośrednie zastosowanie metod do szukania wszystkich reduktów nie ma sensu, gdyż przestrzeń n -elementowych permutacji jest większa niż przestrzeń możliwych podzbiorów atrybutów.

Metody znajdują zastosowanie w zaawansowanych algorytmach wyznaczania minimalnych reduktów. Po pierwsze zastosowane dla całego zbioru atrybutów dostarczają górnego ograniczenia na rozmiar minimalnego reduktu. Dodatkowo metoda druga jest używana do

redukcji otrzymanych w czasie działania algorytmu nadreduktów. W większości zastosowań procedury zachłanne działają dla przyjętej arbitralnie stałej permutacji atrybutów.

6.4.3. Podejście heurystyczne

Duża przestrzeń rozwiązań dla problemu wyznaczania reduktów tablicy decyzyjnej wymusza użycie metod przybliżonych.

Dla problemu wyznaczenia pojedynczego reduktu o małym rozmiarze został zaproponowany algorytm Johnsona. Redukt taki może zostać przykładowo wykorzystany do zbudowania zbioru krótkich reguł klasyfikujących.

Poprawienie wyników metody można osiągnąć przez wstępną optymalizację funkcji rozróżnialności za pomocą praw pochłaniania. Proces ten jest jednak czasochłonny i czyni algorytm nieefektywnym dla dużych danych.

Dla problemu wyznaczania minimalnego reduktu zaproponowano wiele algorytmów przybliżonych. W większości wykorzystują one znane podejścia optymalizacyjne, w szczególności symulowane wyżarzanie lub algorytmy ewolucyjne [49]. Bardzo często istotnymi elementami tych metod są omówione w poprzednim rozdziale procedury zachłanne. Analogiczne podejście optymalizacyjne może zostać zastosowane do rozwiązania interesującego nas problemu wyznaczania wszystkich reduktów tablicy decyzyjnej.

6.4.4. Dekompozycja przestrzeni

Wiele przybliżonych metod wyznaczania reduktów poszukuje zbiorów rozróżniających w całej przestrzeni możliwych podzbiorów atrybutów warunkowych, a następnie redukuje je do reduktów wykorzystując metodę zachłanną. Niestety, dla dużych tablic decyzyjnych rozmiar przestrzeni poszukiwań nie pozwala na użycie algorytmu optymalizacji bezpośrednio do całej przestrzeni.

Dobre efekty daje strategia dekompozycji przestrzeni poszukiwań zaproponowana dla problemu znajdowania minimalnego kolorowania dla grafu [17] oraz problemu wyznaczania minimalnego reduktu [45].

Skupmy się na drugim z wymienionych problemów. Zamiast przeszukiwać bezpośrednio przestrzeń $S=2^C$, możemy rozwiązać co najwyżej n podproblemów dotyczących przeszukiwania podprzestrzeni $S_k = \{X \in 2^C : |X| = k\}$. W przypadku wspomnianych prac zostały zastosowane efektywne heurystyki dla znalezienia podprzestrzeni zawierającej rozwiązanie optymalne. Zauważmy, że jeśli ograniczenie górne znajduje się w podprzestrzeni S_k , to na pewno rozwiązanie całego problemu nie znajduje się w przestrzeni S_m , gdzie $m > k$. W ten sposób można skutecznie zawęzić przestrzeń rozwiązań całego problemu.

Istotna różnica w stosunku do rozważanego problemu wyznaczania wszystkich reduktów polega na tym, że nie zawiera on w sobie problemu optymalizacji i elementy rozwiązania mogą znajdować się w każdej z podprzestrzeni. Mimo to dekompozycja przestrzeni S pozwala na skuteczniejsze przeglądanie podprzestrzeni w poszukiwaniu reduktów.

Proponowany tutaj algorytm wykorzystuje dekompozycję przestrzeni na podprzestrzenie i formułuje problem wyznaczania reduktów za pomocą podproblemów odpowiadających tym podprzestrzeniom. Podproblem k -ty dotyczy wyznaczania minimalnych k -zbiorów trafieniowych.

Dla każdego zbioru z wyznaczonej rodziny k -zbiorów trafieniowych jest znajdowany redukt. Procedura ta ma charakter zachłanny i polega na usuwaniu kolejnych atrybutów zbędnych zgodnie z ustalonym arbitralnie porządkiem. Otrzymane redukty dodawane są do zbioru gromadzonych reduktów.

Poniższy algorytm wyznacza redukty tablicy decyzyjnej o długości nie większej niż *UPPER*. Jeśli przyjmiemy *UPPER*=*n*, to algorytm szuka wszystkich reduktów.

```
funkcja evolutionary-reduct-finder
k=1
LR0=∅
dopóki k≤UPPER
    Hk←wyznacz k-zbiory trafieniowe
    Rk←wyznacz redukty na podstawie Hi
    LRk= LRk-1∪Rk
zwróć LRUPPER
```

6.5. Metody wyszukiwania minimalnych k-zbiorów trafieniowych

6.5.1. Wprowadzenie

Prezentowane tutaj podejścia do problemu wyszukiwania minimalnych *k*-zbiorów trafieniowych polegają na przeszukaniu przestrzeni podzbiorów *k*-elementowych. Formalnie przestrzeń poszukiwań zawiera $\binom{n}{k}$ atrybutów. Jeśli ma ona duże rozmiary, to możliwe jest wyłącznie przejrzanie przybliżone. W ogólności można zaproponować tutaj dowolny algorytm optymalizacji. Poniżej prezentowane są przykładowe metody, które dają kontekst dla dyskusji efektywności.

6.5.2. Podejście dokładne

Metoda sprowadza się do przejrzania całej przestrzeni podzbiorów *k*-elementowych i sprawdzaniu, które są minimalnymi zbiorami trafieniowymi. Ze względu na wyczerpujący charakter działania może być zastosowana wyłącznie dla wartości *k* bliskich 1 lub *n*.

6.5.3. Podejście losowe

Metoda wybiera losowe *k*-elementowe kombinacje i sprawdza, czy są minimalnymi zbiorami trafieniowymi. Może być stosowana dla dowolnych *k*.

```
funkcja random-hitting-set-finder
LH=∅
dopóki (nie warunek stopu)
    H←wybierz losowo zbiór
    jeśli H jest minimalnym zbiorem trafieniowym to
        LH= LH∪H
zwróć LH
```

6.5.4. Podejście ewolucyjne

Znacznie skuteczniejsze okazują się metody, które odwołują się w danej chwili do historii przeprowadzonych obliczeń oraz wybierają obszary poszukiwań odwołując się do miar przydatności rozwiązań. Przykładem takiego podejścia są algorytmy ewolucyjne [3,17].

Algorytmy ewolucyjne odwołują się do teorii ewolucji. Z tego względu do opisu niektórych pojęć informatycznych powszechnie stosowane jest słownictwo zaczerpnięte z genetyki. Nie należy przy tym interpretować tych pojęć dosłownie, gdyż ich obecność związana jest

wyłącznie z istnieniem pewnej analogii pomiędzy proponowaną metodą poszukiwań, a procesami biologicznymi.

Przestrzeń poszukiwań jest mapowana na przestrzeń osobników. Z każdym osobnikiem związany jest jego genotyp. Wprowadza się wzajemnie jednoznaczne kodowanie rozwiązań na genotypy. Stopień przystosowania osobnika do środowiska mierzy jakość rozwiązania, tak więc przystosowanie jest bezpośrednio związane z optymalizowaną w zadaniu funkcją.

Algorytm rozpatruje pewną populację osobników w określonym horyzoncie czasowym. Wybór osobników do populacji początkowej jest określany, jako etap inicjacji lub siania. Jednostka czasu jest związana ze zmianą jednego pokolenia (generacji), czyli zastąpieniu starej populacji, zwanej bazową, przez następną. Przemiana pokoleń realizowana jest poprzez symulowaną ewolucję. Nowa populacja powstaje ze starej przez reprodukcję osobników i sekwencyjne zastosowanie określonego łańcucha operatorów genetycznych.

Reprodukcja związana jest z odpowiednią selekcją osobników, które są umieszczane w populacji tymczasowej. Selekcja jest dyktowana stopniem przystosowania osobników i musi być poprzedzona ocenieniem przystosowania w całej populacji.

Pojedynczy operator działa na wybranych osobnikach, zwanych dalej rodzicielskimi, i daje w wyniku nowe osobniki, zwane dalej potomnymi. Operator określa zależność pomiędzy genotypami osobników potomnych i rodzicielskich.

Tradycyjnie łańcuch operatorów składa się z operatorów krzyżowania i mutacji. Przy wielu powszechnie stosowanych kodowaniach zachowują one analogię do odpowiednich zjawisk genetycznych. Krzyżowanie przekształca dwa osobniki rodzicielskie na dwa osobniki potomne. Mutacja zaś na podstawie jednego rodzica tworzy jednego osobnika potomnego.

Otrzymana populacja wynikowa i populacja bazowa dają w procesie sukcesji populację następnej generacji.

Czas działania algorytmu określony jest przez warunek stopu. Może być związany zarówno z liczbą generacji, bieżącą populacją lub wielkością określoną dla całej symulacji.

Rozwiązanie otrzymuje się poprzez badanie populacji dla kolejnych generacji i rejestrację określonej ilości osobników rekordowych.

W zarysie schematu algorytmu można zapisać następująco [3]:

```
funkcja evolutionary-algorithm
t ← 0
inicjacja  $P^0$ 
ocena  $P^0$ 
 $R^0$  ← rejestruj najlepsze wyniki z  $P^0$ 
dopóki (nie warunek stopu)
     $T^t$  ← reprodukcja  $P^t$ 
     $O^t$  ← zastosowanie operatorów na  $T^t$ 
    ocena  $O^t$ 
     $P^{t+1}$  ← sukcesja z  $P^t$  i  $O^t$ 
     $R^{t+1}$  ← rejestruj najlepsze wyniki z  $P^{t+1}$ 
    t ← t + 1
zwróć  $R^t$ 
```

Techniki ewolucyjne znajdują szerokie zastosowanie w problemach optymalizacji. Tym niemniej, nie został dla nich dotąd sformułowany dowód zbieżności. Za nieformalne uzasadnienie przyjmuje się obserwację, że natura nie musi działać optymalnie, ale mimo to wydaje się działać dobrze.

Rozważając problem poszukiwania k -zbiorów trafieniowych musimy wpierv zaproponować mapowanie przestrzeni poszukiwań na przestrzeń osobników. Pojedynczy osobnik reprezentuje k -zbiór trafieniowy A . Genotyp osobnika jest wektorem n -elementowym P , takim że:

$$P_i = \begin{cases} 1 & a_i \in A \\ 0 & a_i \notin A \end{cases}$$

Powyżej przyjęte kodowanie zachowuje stałą liczbę jedynek w wektorze P , równą k .

Początkowo populacja inicjowana jest osobnikami o genotypach losowanych z równomiernego rozkładu kombinacji k -elementowych.

Operatory genetyczne muszą zachowywać niezmienną liczbę jedynek w genotypach, dlatego nie możliwe jest zastosowanie standardowych operatorów dla kodowania binarnego.

Wygodne jest zastosowanie kolejno dwóch operatorów. Każdy z operatorów jest stosowany z określonym prawdopodobieństwem.

Operator krzyżowania działa na 2 osobnikach rodzicielskich i zwraca 2 osobniki potomne. Analizowane są kolejne geny obu osobników, zgodnie z poniższym algorytmem.

Oznaczenia:

A, B – genotypy osobników rodzicielskich

C, D – genotypy osobników potomnych

$$ones(A, k) = \sum_{i=1}^{k-1} A_i$$

funkcja cross-over (A, B)

dla każdego i **od** 1 **do** n

jeśli $A_i = B_i$, **to**

$C_i \leftarrow A_i$

$D_i \leftarrow A_i$

inaczej jeśli $ones(A, i) < ones(B, i)$ **to**

$C_i \leftarrow 1$

$D_i \leftarrow 0$

inaczej jeśli $ones(A, i) > ones(B, i)$ **to**

$C_i \leftarrow 0$

$D_i \leftarrow 1$

Inaczej

$X \leftarrow \text{losuj spośród zbioru } \{C, D\}$

$Y \leftarrow \text{element zbioru } \{C, D\} \setminus \{X\}$

$X_i \leftarrow 1$

$Y_i \leftarrow 0$

Operator mutacji działa na i i daje w wyniku jednego osobnika. Polega on na zamianie miejscami dwóch losowo wybranych genów.

Funkcja przystosowania musi być tak zdefiniowana, aby proponować osobniki reprezentujące zbiory o większej sile rozróżniania obiektów. Siłę tę możemy rozumieć, jako liczbę par obiektów pokrytych przez zbiór reprezentowany danym osobnikiem. Ze względu na wygodę odnosimy tę miarę do siły rozróżniającej całego zbioru atrybutów, dzięki czemu zbiór wartości funkcji przystosowania zawiera się w przedziale $<0, 1>$. Funkcja przystosowania przyjmuje postać:

$$fitness(B) = \frac{|CS(B)|}{|CS(A)|}$$

Zadanie optymalizacji poszukuje rozwiązania maksymalizującego tę funkcję celu.

7. Wyznaczanie charakterystycznych wzorców ekspresji

7.1. Wprowadzenie

Jednym z celów stawianych przed dziedziną odkrywania wiedzy jest analiza danych pod kątem wykrywania pewnych charakterystycznych wzorców. W pracy zajmę się szczególnym typem wzorców – wzorcami wyłaniającymi.

Idea wzorców wyłaniających jest stosunkowo młoda. Pierwsze kroki w tym temacie przypisuje się Li, a pierwszą definicję tej nowej formy zapisu wiedzy można znaleźć w publikacji [13] z 1999. Pojęcie związane jest z badaniem pary dyskretnych systemów informacyjnych o atrybutach z tymi samymi przeciwdziedzinami. Wzorzec jest zbiorem par (atrybut, wartość atrybutu). Istotność wzorca zależy od tego, jak bardzo jest on charakterystyczny dla jednego z systemów. Precyzując, wzorzec jest istotny, jeśli dla niewielkiej części obiektów pierwszego systemu atrybuty przyjmują jednocześnie wskazywane przez wzorzec wartości, podczas gdy dla drugiego systemu własność ta zachodzi dla relatywnie większej części obiektów. Mówi się, że wzorzec występuje z systemu pierwszego do drugiego. Stosunek częstotliwości występowania wzorca wyraża jego siłę rozróżniania pomiędzy systemami i jest określany mianem współczynnika wzrostu wsparcia. Najwięcej miejsca poświęciłem wzorcom wyskakującym, które w ogóle nie występują w jednym systemie, ale są obecne w drugim. Mają one nieskończony współczynnik wzrostu wsparcia.

Autorzy wskazują, że wzorce mogą być wykorzystane do analizy trendów w bazach danych znakowanych czasem oraz do znajdowania kontrastów między klasami.

Rozważmy tablicę decyzyjną, w której:

- każdy obiekt dotyczy pojedynczego zakupionego koszyka
- każdy atrybut warunkowy jest binarny i odpowiada zakupowi pewnego artykułu w sklepie
- atrybut decyzyjny określa rok zarejestrowania koszyków i przyjmuje dwie wartości ze zbioru {2000,2001}

Wzorce pozwalają na identyfikację zestawów artykułów, które były często kupowane w jednym roku i rzadko w drugim. Przykładowy wzorzec wyłaniający z roku 2000 do 2001: {(żurek,1),(chleb,1),(kiełbasa biała,1)}. Spadek występowania zestawu w roku następnym mógł wynikać z upowszechnienia informacji o szkodliwych składnikach wykrytych w zupie żurek pochodzącej od największych producentów.

W pracy [28] rozważono popularny zestaw danych na temat grzybów jadalnych i trujących [58]. Znalaziono dla niego miliony wzorców o współczynniku wzrostu wsparcia nie mniejszym niż 2,5. Przykładowy wzorzec wyskakujący ze zbioru grzybów jadalnych do grzybów trujących: {(Odor,none),(Gill_Size,broad),(Ring_Number,one)}.

Wzorce wyłaniające posiadają istotne własności [27]. Po pierwsze badania wykazały, że są one poprawnym zapisem wiedzy. Jeśli obiekty odpowiadają pomiarom rzeczywistych pojęć, to wzorce wyznaczone dla pewnego zbioru obiektów pozostają z dużym prawdopodobieństwem wzorcami w zbiorze rozszerzonym. Fakt ten wskazuje na dobre zachowanie wzorców przy opisie natury danego zagadnienia. Zauważmy dalej, że tradycyjne metody statystyczne nie są w stanie wykryć wzorców o dużej długości. Wzorce takie stanowią zupełnie nowe źródło informacji. Co więcej postać wzorca jest łatwa do zrozumienia

dla człowieka, gdyż można ją interpretować, jako koniunkcję pewnych warunków i bezpośrednio przełożyć na regułę.

Zgodnie z przykładami przytoczonymi wcześniej wzorce znajdują szerokie zastosowanie praktyczne. Autorzy pomysłu szczególnie w wielu pracach omówili znakomite zdolności wzorców w zadaniach klasyfikacji. Wymienić tu należy leniwy klasyfikator DeEP [27], klasyfikator CAEP [14] oraz JEP-classifier [28]. Dla większości analizowanych zbiorów danych zaproponowane klasyfikatory wykazały się najwyższą skutecznością w porównaniu do innych popularnych metod (m.in. CBA, LB, C5.0, k-NN). W zależności od sposobu zdefiniowania problemu wzorcami można posłużyć się zarówno przy znajdowaniu różnic pomiędzy dwoma klasami, jak i do predykcji trendów czasowych.

Narzuca się również wykorzystanie wzorców do znajdowania zbiorów reguł asocjacyjnych [27]. W szczególności możliwe jest wyznaczenie reguł o niskim wsparciu, które trudno wykryć za pomocą algorytmów tradycyjnych, bazujących na znajdowaniu zbiorów częstych dla określonego progu wsparcia. Ponadto, wzorce wyskakujące mogą być użyte do formułowania zbiorów reguł pewnych. Duże znaczenie odgrywają tutaj wzorce o dużym wsparciu, czyli takie, które są szczególnie charakterystyczne dla obiektów jednej klasy.

7.2. Znaczenie dla badania ekspresji

Pierwotnie idea wzorców wyłaniających została opracowana do analizy danych biznesowych. Po pierwszych sukcesach zakres zastosowań powoli zaczął się rozszerzać. Wykorzystanie wzorców do badania ekspresji genów zaproponowali autorzy pojęcia [30,31].

Metoda wymaga wstępnego oszacowania istotności genów i eliminacji najmniej znaczących ze względów efektywnościowych, a następnie dyskretyzacji.

Podejście zmierza do identyfikacji wzorców wyłaniających. W tym przypadku są to zbiory par (gen, zbiór wartości), które są charakterystyczne dla obiektów jednej klasy decyzyjnej i rzadkie dla obiektów innych klas. Przykładowy wzorzec może mieć postać:

$$\{(K03001, <89.20, +\infty), (R76254, <127.16, +\infty), (D31767, <63.03, +\infty))\}.$$

Wzorzec ten pokazuje, że trzy geny ulegają razem ekspresji w większości tkanek danego typu. Przedziały wartości pochodzą z kroku dyskretyzacji.

Wzorce dostarczają informacji o grupach funkcjonalnych genów, które pozostają w silnym związku z badanym zjawiskiem. Mogą one być również z powodzeniem wykorzystane do budowy skutecznych klasyfikatorów tkanek wnioskujących na podstawie profilu ekspresji.

W naszej analizie wykorzystamy podejście do badania ekspresji zaproponowane przez autorów idei. Rozważając zdyskretyzowaną macierz ekspresji będziemy poszukiwać najczęściej występujących wzorców wyskakujących, które wskazują najistotniejsze grupy genów.

7.3. Definicje i twierdzenia

Poniżej wprowadzam pojęcia powszechnie wykorzystywane w zagadnieniach reprezentacji wiedzy. Charakter pojęć odwołuje się do modelu zakupów w sklepie, w którym można nabyć po jednym przedmiocie każdego rodzaju [20]. Zawartość pojedynczego koszyka jest rejestrowana pod postacią transakcji w prostej jednotabelowej bazie danych. Transakcja jest tutaj rozumiana, jako podzbiór przedmiotów wybranych z pewnej ustalonej przestrzeni. Na

gruncie teorii relacyjnych baz danych powyższa baza danych może być interpretowana jako relacja, dla której każdy atrybut odpowiada wzajemnie jednoznacznie przedmiotowi oraz domena każdego atrybutu jest binarna. Arbitralnie możemy przyjąć, że wartość 1 danego atrybutu dla pewnej krotki oznacza obecność odpowiadającego mu przedmiotu w transakcji opisywanej tą krotką.

Ogólnie cel powyższego rozumowania jest dyktowany wygodą i sprowadza się do skojarzenia pojedynczego obiektu systemu informacyjnego ze zbiorem wartości, który jednoznacznie opisuje wartości wszystkich atrybutów dla rozważanego obiektu.

Wszystkie definicje dotyczące budowy rodziny baz danych dla tablicy decyzyjnej zostały zaproponowane w celu połączenia zagadnień opartych na reprezentacji informacji w postaci tablicy decyzyjnej i baz danych. Definicje i część twierdzeń dotyczących wzorców wyłaniających pochodzi z pracy [27], przy czym zostały one przeformułowane do przyjętej w rozdziale konwencji. Twierdzenia wprowadzane na potrzeby pracy zostały poparte dowodami.

Rozważmy system informacyjny (U, A) o skończonych atrybutach dyskretnych. Założenie to dotyczy wszystkich rozważanych w tym rozdziale systemów informacyjnych.

Definicja 7.3-1

Przedmiot oznacza parę (a, v) , gdzie $a \in A$ i $v \in V_a$.

Definicja 7.3-2

Niech $IS = (U, A)$ będzie systemem informacyjnym.

Przestrzeń przedmiotów S_{IS} oznacza zbiór przedmiotów $S_{IS} = \{(a, v) : a \in A \wedge v \in V_a\}$.

Przykład

Dla przybliżenia pojęć będziemy w tym rozdziale rozważać tablicę decyzyjną $DT_3 = (U, A, C, \{d\})$ otrzymaną w rozdziale 5.3.1. Rozpatrzmy system informacyjny $IS = (\{H_1, H_2, H_3, H_4\}, C)$, który zawiera obiekty $u \in U$, takie że $d(u) = 0$.

	G ₁	G ₂	G ₃	G ₄	G ₅
H ₁	0	0	0	1	1
H ₂	0	0	0	1	1
H ₃	0	0	1	0	0
H ₄	0	0	1	1	1

Tab. 7.3-1. Tablica decyzyjna DT_3

Przestrzeń przedmiotów $S_{IS} = \{(G_1, 0), (G_2, 0), (G_2, 1), (G_3, 0), (G_3, 1), (G_4, 0), (G_4, 1), (G_5, 0), (G_5, 1)\}$.

Definicja 7.3-3

Niech $IS = (U, A)$ będzie systemem informacyjnym.

Przyporządkowanie transakcji T_{IS} dla systemu informacyjnego IS oznacza bijekcję $T_{IS} : U \rightarrow 2^{S_{IS}}$, taką że $T_{IS}(u) = \{(a, v) \in A \times V_a : a \in A \wedge a(u) = v\}$.

Przy czym 2^A oznacza zbiór potęgowy dla zbioru A .

Definicja 7.3-4

Niech $IS=(U,A)$ będzie systemem informacyjnym, T_{IS} jego przyporządkowaniem transakcji, a $u \in U$ przedmiotem.

Transakcja dla danego obiektu u oznacza zbiór przedmiotów $T_{IS}(u)$.

Przestrzeń przedmiotów jest zbiorem wszystkich przedmiotów, które mogą pojawić się w transakcjach.

Przykład

Przyporządkowanie transakcji dla systemu informacyjnego IS przedstawione zostało w tabeli 7.3-2. Wartość przyporządkowania dla każdego obiektu jest transakcją.

u	$T_{IS}(u)$
H_1	$\{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}$
H_2	$\{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}$
H_3	$\{(G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0)\}$
H_4	$\{(G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1)\}$

Tab. 7.3-2. Przyporządkowanie transakcji T_{IS}

Definicja 7.3-5

Niech $IS= (U,A)$ będzie systemem informacyjnym.

Baza danych D_{IS} dla systemu informacyjnego IS oznacza multizbiór transakcji $D_{IS} = \{T_{IS}(u) : u \in U\}$.

Przyporządkowanie transakcji wyraża wzajemną odpowiedniość pomiędzy systemem informacyjnym i jego bazą danych. Ze względu na tę odpowiedniość bazę danych często rozpatruje się w oderwaniu od systemu informacyjnego. Jeśli nie będzie to prowadziło do niejednoznaczności, to w oznaczeniu bazy danych *implicite* będzie przyjmowane, że dotyczy ona określonego systemu informacyjnego i indeks dolny zostanie pominięty.

Przykład

Baza danych dla systemu IS jest multizbiorem następującej postaci:

$$\{ \begin{array}{l} \{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}, \\ \{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}, \\ \{(G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0)\}, \\ \{(G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1)\} \end{array} \}$$

Baza została przedstawiona przez enumerację swoich elementów. Należy pamiętać, że formalnie jest ona funkcją, która wyraża liczbę wystąpień danego zbioru (punktu dziedziny) w multizbiorze. Rozważany multizbiór ma postać:

X	$N(x)$
$\{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}$	2
$\{(G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0)\}$	1
$\{(G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1)\}$	1

Tab. 7.3-3. Baza danych D_{IS}

Otrzymana baza danych będzie dalej oznaczana przez D_I .

Definicja 7.3-6

Niech D będzie bazą danych.

Liczność występowania zbioru przedmiotów X w bazie danych D oznacza $count_D(X) = |\{u \in D : X \subset D\}|$.

Definicja 7.3-7

Niech D będzie bazą danych.

Wsparcie zbioru przedmiotów X w bazie danych D oznacza $supp_D(X) = \frac{count_D(X)}{|D|}$.

Definicja 7.3-8

Niech D będzie bazą danych, a δ ustaloną liczbą rzeczywistą, przy czym $\delta \in (0,1)$.

Zbiór przedmiotów X w bazie danych D dla ustalonego poziomu wsparcia δ jest:

- **częsty**, jeśli $supp_D(X) \geq \delta$
- **rzadki**, jeśli $supp_D(X) < \delta$

Przykład

W tabeli 7.3-4 zostały przedstawione przykładowe zbiory przedmiotów, dla których wyznaczono licznosc i wsparcie oraz określono, czy zbiór jest częsty czy rzadki dla $\delta=50\%$.

Liczność przedmiotów X	$count(X)$	$supp(X)$	częsty/rzadki
$\{(G_1,0),(G_2,0)\}$	4	100%	częsty
$\{(G_2,0),(G_4,1)\}$	3	75%	częsty
$\{(G_2,0),(G_5,1)\}$	3	75%	częsty
$\{(G_3,1),(G_4,1),(G_5,1)\}$	1	25%	rzadki

Tab. 7.3-4. Licznosc i wsparcie zbiorów przedmiotów w D_{IS}

Rozważmy tablicę decyzyjną $(U,A,C,\{d\})$ o skończonych i dyskretnych atrybutach warunkowych oraz skończonym atrybucie decyzyjnym d . Założenie to dotyczy wszystkich rozważanych w tym rozdziale tablic decyzyjnych.

Definicja 7.3-9

Relacja przynależności do klasy R_{DT} dla tablicy decyzyjnej DT , jest relacją dwuargumentową w zbiorze U o postaci $R_{DT} = \{(u_1, u_2) \in U^2 : d(u_1) = d(u_2)\}$.

Twierdzenie 7.3-1

Relacja przynależności do klasy dla tablicy decyzyjnej DT jest relacją równoważności.

Dowód:

Dla równoważności relacji przynależności do klasy potrzeba i wystarcza jej zwrotność, symetryczność i przechodniość. Cechy te przenoszą się z relacji równości w zbiorze V_d .

Weźmy dowolne $u \in U$. Relacja jest zwrotna, gdyż zachodzi $d(u) = d(u)$.

Weźmy dowolne $u, v \in U$. Relacja jest symetryczna, gdyż $d(u) = d(v) \Rightarrow d(v) = d(u)$.

Weźmy dowolne $u, v, w \in U$. Relacja jest przechodnia, gdyż

$$d(u) = d(v) \wedge d(v) = d(w) \Rightarrow d(u) = d(w).$$

□

Dla relacji R w zbiorze A przyjmuję dalej oznaczenia:

- $[a]_R$ oznacza klasę abstrakcji relacji R generowaną przez element $a \in A$
- $A|_R$ oznacza zbiór ilorazowy relacji R

Relacja przynależności do klasy R_{DT} generuje podział zbioru U na klasy abstrakcji. Zauważmy, że każda klasa abstrakcji związana jest wzajemnie jednoznacznie z pewną wartością atrybutu decyzyjnego. Klasy abstrakcji są niepuste, dlatego istnieje przynajmniej jeden obiekt należący do każdej z nich.

Oznaczmy:

- $V_d = \{d_1, \dots, d_p\}$
- dla każdego $i = 1, \dots, p$: $U_i = [u_i]_{R_{DT}}$ przy dowolnym obiekcie u_i takim, że $d(u_i) = d_i$

Przykład

W celu przybliżenia kolejnych twierdzeń zostanie wykorzystana tablica decyzyjna DT_3 .

	G ₁	G ₂	G ₃	G ₄	G ₅	d
H ₁	0	0	0	1	1	Z
H ₂	0	0	0	1	1	Z
H ₃	0	0	1	0	0	Z
H ₄	0	0	1	1	1	Z
H ₅	0	1	0	0	0	R
H ₆	0	1	0	0	0	R
H ₇	1	1	0	1	0	R
H ₈	1	1	0	0	0	R

Tab. 7.3-5. Tablica decyzyjna DT_3

Relacja przynależności do klasy ma postać:

$$R_{DT} = \{ (H_1, H_1), (H_1, H_2), (H_1, H_3), (H_1, H_4), (H_2, H_1), (H_2, H_2), (H_2, H_3), (H_2, H_4), \\ (H_3, H_1), (H_3, H_2), (H_3, H_3), (H_3, H_4), (H_4, H_1), (H_4, H_2), (H_4, H_3), (H_4, H_4), \\ (H_5, H_5), (H_5, H_6), (H_5, H_7), (H_5, H_8), (H_6, H_5), (H_6, H_6), (H_6, H_7), (H_6, H_8), \\ (H_7, H_5), (H_7, H_6), (H_7, H_7), (H_7, H_8), (H_8, H_5), (H_8, H_6), (H_8, H_7), (H_8, H_8) \}$$

Zbiór ilorazowy relacji przynależności R_{DT} ma postać:

$$A|_R = \{\{H_1, H_2, H_3, H_4\}, \{H_5, H_6, H_7, H_8\}\}$$

Definicja 7.3-10

Niech $DT=(U,A,C,\{d\})$ będzie tablicą decyzyjną i R_{DT} relacją przynależności.

Rodzina systemów informacyjnych ze względu na decyzję dla tablicy decyzyjnej DT oznacza zbiór systemów informacyjnych $\{(U_1, C_1), \dots, (U_p, C_p)\}$, przy czym:

- $\{U_1, \dots, U_p\} = U|_{R_{DT}}$
- dla każdego $i=1, \dots, p$ zachodzi $C_i = \{a_{i1}, \dots, a_{in}\}$, przy czym dla każdego $j=1, \dots, n$ mamy $a_{ij} : U_i \rightarrow V_j$

Definicja 7.3-11

Niech $DT=(U,A,C,\{d\})$ będzie tablicą decyzyjną i $\{(U_1, C_1), \dots, (U_p, C_p)\}$ będzie rodziną systemów informacyjnych ze względu na decyzję dla tej tablicy.

Rodzina baz danych ze względu na decyzję DBF dla tablicy decyzyjnej DT oznacza zbiór baz danych $\{D_1, \dots, D_p\}$, taki że dla każdego $i=1, \dots, p$: D_i jest bazą danych dla systemu informacyjnego (U_i, C_i) .

Dla jasności zapisu warto podsumować zachodzące odpowiedniości pomiędzy omówionymi pojęciami. Rozważając tablicę decyzyjną $DT=(U,A,C,\{d\})$ wprowadzamy arbitralną numerację wartości, które przyjmuje atrybut decyzyjny d . Każdej wartości d_i tego atrybutu odpowiada wzajemnie jednoznacznie klasa abstrakcji relacji przynależności do klasy R_{DT} . Dla każdej klasy abstrakcji możemy zdefiniować system informacyjny tworząc w ten sposób rodzinę systemów informacyjnych ze względu na decyzję. Ponieważ nasze dalsze rozważania operować będą na bazach danych, dlatego każdy system informacyjny rodziny kojarzymy z odpowiadającą mu bazą danych i otrzymujemy rodzinę baz danych ze względu na decyzję. Zauważmy, że ze względu na pierwotną tablicę decyzyjną DT wszystkie bazy danych rodziny mają wspólną przestrzeń przedmiotów.

Stosunkowo często mamy do czynienia z przypadkiem dwóch klas decyzyjnych, który zachodzi dla $p=2$. W tym przypadku rodzina baz danych ze względu na decyzję jest zbiorem dwuelementowym $\{D_1, D_2\}$. Można również sztucznie wyróżnić jedną interesującą nas klasę, a wszystkie inne potraktować jako klasę przeciwną. W ten sposób atrybut decyzyjny d zastępujemy nowym atrybutem d' o dwuelementowym zbiorze wartości.

Przykład

Rodzina baz danych ze względu na decyzję $DBF_4=\{D_1, D_2\}$ dla tablicy decyzyjnej DT_3 ma postać:

$$\left\{ \begin{array}{l} \{ (G_1,0), (G_2,0), (G_3,0), (G_4,1), (G_5,1) \}, \\ \{ (G_1,0), (G_2,0), (G_3,0), (G_4,1), (G_5,1) \}, \\ \{ (G_1,0), (G_2,0), (G_3,1), (G_4,0), (G_5,0) \}, \\ \{ (G_1,0), (G_2,0), (G_3,1), (G_4,1), (G_5,1) \} \end{array} \right\},$$

$$\left\{ \begin{array}{l} \{ (G_1,0), (G_2,1), (G_3,0), (G_4,0), (G_5,0) \}, \\ \{ (G_1,0), (G_2,1), (G_3,0), (G_4,0), (G_5,0) \}, \\ \{ (G_1,1), (G_2,1), (G_3,0), (G_4,1), (G_5,0) \}, \\ \{ (G_1,1), (G_2,1), (G_3,0), (G_4,0), (G_5,0) \} \end{array} \right\}$$

Rodzina będzie dalej wykorzystywana do dalszych obliczeń. Indeks rodziny został tak dobrany, aby zobrazować kolejny etap analizy.

Definicja 7.3-12

Niech $\{D_1, D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy decyzyjnej DT .
Współczynnik wzrostu wsparcia z D_1 do D_2 dla zbioru przedmiotów X oznacza

$$growth_{D_1 \rightarrow D_2}(X) = \begin{cases} 0 & \text{dla } supp_1(X) = 0 \text{ i } supp_2(X) = 0 \\ \infty & \text{dla } supp_1(X) = 0 \text{ i } supp_2(X) \neq 0 \\ \frac{supp_2(X)}{supp_1(X)} & \text{w innym przypadku} \end{cases}$$

Definicja 7.3-13

Niech $\{D_1, D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy decyzyjnej DT , a ρ liczbą rzeczywistą, dla której zachodzi $\rho > 1$.

ρ -wzorzec wyłaniający z D_1 do D_2 oznacza zbiór przedmiotów X , dla którego $growth_{D_1 \rightarrow D_2}(X) \geq \rho$.

Ze względu na nomenklaturę angielską ρ -wzorzec wyłaniający określa się w skrócie ρ -EP (ρ -emerging pattern). Jeśli w danym kontekście ρ jest jawnie ustalone, to możemy mówić po prostu o wzorcu wyłaniającym (EP). Ponadto, jeśli baza odniesienia D_1 jest ustalona, powszechnie używane jest również określenie wzorzec wyskakujący w D_2 .

Definicja 7.3-14

Niech $\{D_1, D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy decyzyjnej DT .

Wzorzec wyskakujący z D_1 do D_2 oznacza wzorzec wyłaniający, dla którego $growth_{D_1 \rightarrow D_2}(X) = \infty$.

Wzorzec wyskakujący odpowiada angielskiemu jumping emerging pattern. Dalej używany będzie skót JEP.

Twierdzenie 7.3-2

Wzorzec wyłaniający X jest wyskakujący z D_1 do D_2 , wtedy i tylko wtedy, gdy $supp_1(X) = 0 \wedge supp_2(X) \neq 0$.

Dowód:

Trywialny, wynika z definicji 7.3-12 i 7.3-14. □

Przykład

W tabeli 7.3-6 zamieszczone są przykłady wzorców wyskakujących. Zostały wybrane największe wartości parametru ρ , takie że zbiór X jest ρ -wzorcem wyłaniającym. Dla wzorców wyskakujących wsparcie w bazie, z której są wyznaczane pozwala na ocenę, jak często występują w transakcjach związanych z daną klasą.

Zbiór przedmiotów X	$supp_1(X)$	$supp_2(X)$	$\sigma = growth_{D_1 \rightarrow D_2}(X)$	Typ wzorca
$\{(G_1,0)\}$	100%	50%	2	wyłaniający
$\{(G_4,1)\}$	75%	25%	3	wyłaniający
$\{(G_1,0),(G_4,1)\}$	75%	0	∞	wyskakujący
$\{(G_1,0),(G_2,0),(G_3,0),$ $(G_4,1),(G_5,1)\}$	50%	0	∞	wyskakujący

Tab. 7.3-6. Przykłady wzorców wyłaniających

Definicja 7.3-15

Niech $DBF=\{D_1,D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy decyzyjnej DT , a ρ liczbą rzeczywistą, dla której zachodzi $\rho > 1$.

Przestrzeń ρ -wzorców wyłaniających z D_1 do D_2 oznacza zbiór wszystkich ρ -wzorców wyłaniających z D_1 do D_2 .

Przestrzeń ρ -wzorców wyłaniających z D_1 do D_2 oznaczamy przez $\rho\text{-}EP_{DBF}(D_1,D_2)$.

Definicja 7.3-16

Niech $DBF=\{D_1,D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy DT .

Przestrzeń wzorców wyskakujących z D_1 do D_2 oznacza zbiór wszystkich wzorców wyskakujących z D_1 do D_2 .

Przestrzeń ρ -wzorców wyłaniających z D_1 do D_2 oznaczamy przez $JEP_{DBF}(D_1,D_2)$.

Definicja 7.3-17

Niech S będzie dowolnym zbiorem skończonym.

Kolekcja C podzbiorów S jest **zamknięta ze względu na podzbiory**, jeśli $\forall X \in C \ \forall Y \in 2^C : Y \in C$.

Innymi słowy, kolekcja jest zamknięta ze względu na podzbiory, jeśli każdy podzbiór dowolnego elementu kolekcji również należy do kolekcji.

Dokładna analiza rozważanych pojęć [27] pokazuje, że nie wszystkie przestrzenie ρ -wzorców wyłaniających są zamknięte ze względu na podzbiory. Stąd, podzbiór właściwy ρ -wzorca wyłaniającego nie musi być ρ -wzorcem wyłaniającym.

Definicja 7.3-18

Niech $\{D_1,D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy decyzyjnej DT , a ρ liczbą rzeczywistą, dla której zachodzi $\rho > 1$.

ρ -wzorec wyłaniający jest **silny**, jeśli wszystkie jego niepuste podzbiory są ρ -wzorcami wyłaniającymi.

Przykład

Niech $S=\{(G_1,0),(G_4,1)\}$. Rozważmy kolekcję $C=\{\{(G_1,0)\},\{(G_4,1)\},\{(G_1,0),(G_4,1)\}\}$. Kolekcja C jest zamknięta ze względu na podzbiory. Zgodnie z powyższym 2-wzorec wyłaniający $\{(G_1,0),(G_4,1)\}$ jest silny, gdyż zbiory przedmiotów $\{(G_1,0)\}$ i $\{(G_4,1)\}$ są również 2-wzorcami wyłaniającymi.

Z kolei 3-wzorec wyłaniający $\{(G_1,0),(G_4,1)\}$ nie jest silny, gdyż $growth_{D_1 \rightarrow D_2}(\{(G_1,0)\}) = 2 < 3$.

Definicja 7.3-19

Niech S będzie dowolnym zbiorem skończonym.

Rodzina C podzbiorów S jest **wypukła**, jeśli $\forall X, Y, Z \subset S : X, Z \in C \wedge X \subset Y \subset Z \Rightarrow Y \in C$.

Przykład

Niech $S = \{a, b, c, d\}$.

Rodzina $\{\{a\}, \{a, c\}, \{a, d\}, \{a, c, d\}\}$ jest wypukła.

Z kolei rodzina $\{\{b\}, \{a, b, c\}\}$ nie jest wypukła, gdyż nie zawiera zbiorów $\{a, b\}$ i $\{b, c\}$.

Definicja 7.3-20

Niech S będzie dowolnym zbiorem skończonym.

Antylańcuch oznacza rodzinę C podzbiorów S , dla którego zachodzi

$\forall (X, Y \in C) : X \neq Y \Rightarrow (X \not\subset Y \wedge Y \not\subset X)$.

Przykład

Niech $S = \{a, b, c, d\}$. Rodzina $\{\{b, d\}, \{a, c\}, \{a, d\}\}$ jest antylańcuchem.

Z kolei rodzina $\{\{a, b, c\}, \{a, c\}, \{a, b, d\}\}$ nie jest antylańcuchem, gdyż $\{a, c\} \subset \{a, b, c\}$.

Definicja 7.3-21

Niech S będzie dowolnym zbiorem skończonym.

Płot oznacza parę (L, R) , dla której:

- L, R są rodzinami podzbiorów S
- L, R są antylańcuchami
- $\forall (X \in L) \exists (Y \in R) : X \subset Y$
- $\forall (Y \in R) \exists (X \in L) : X \subset Y$

Płot oznaczamy przez $\langle L, R \rangle$. Rodzinę L nazywamy lewym krańcem płotu, a rodzinę R prawym krańcem płotu.

Definicja 7.3-22

Niech S będzie dowolnym zbiorem skończonym.

Przedział zbiorów $[L, R]$ reprezentowany przez płot $\langle L, R \rangle$ oznacza rodzinę

$\{Y \subset S : \exists (X \in L) \exists (Z \in R) : X \subset Y \subset Z\}$

Przykład

Niech $S = \{a, b, c, d\}$.

Dla rodzin zbiorów zachodzi:

- $C = \{\{a\}, \{a, c\}, \{a, d\}, \{a, c, d\}\} = [\{\{a\}\}, \{\{a, c, d\}\}]$. Przy czym rodzinę wypukłą C reprezentuje płot $\langle \{\{a\}\}, \{\{a, c, d\}\} \rangle$
- $C = \{\{a, b\}, \{a, c\}, \{a, b, c\}\} = [\{\{a, b\}, \{a, c\}\}, \{\{a, b, c\}\}]$. Przy czym rodzinę wypukłą C reprezentuje płot $\langle \{\{a, b\}, \{a, c\}\}, \{\{a, b, c\}\} \rangle$

Twierdzenie 7.3-3 ([27])

Rodzina jest wypukła wtedy i tylko wtedy, gdy jest przedziałem zbiorów.

Ze względu na pełną odpowiedniość pojęcia rodziny wypukłej i przedziału zbiorów twierdzenia sformułowane dla jednego z nich są prawdziwe także dla drugiego.

Twierdzenie 7.3-4 ([27])

Dla każdej przestrzeni wypukłej C istnieje dokładnie jeden płot $\langle L, R \rangle$, przy czym:

- L jest zbiorem elementów minimalnych C
- R jest zbiorem elementów maksymalnych C

Definicja 7.3-23

Niech D będzie bazą danych dla systemu informacyjnego IS i S jej przestrzenią przedmiotów.

Przestrzeń horyzontalna dla bazy danych D oznacza $\{X \in S : \text{supp}(X) \neq \emptyset\}$.

Przestrzeń horyzontalna jest związana z jedną bazą danych i zawiera wszystkie zbiory przedmiotów, które mają niezerowe wsparcie w tej bazie.

Twierdzenie 7.3-5 ([27])

Niech D będzie bazą danych dla systemu informacyjnego IS .

Przestrzeń horyzontalna dla bazy D jest wypukła.

Przykład

Rodzina baz danych ze względu na decyzję $DBF_4 = \{D_1, D_2\}$ dla tablicy decyzyjnej DT_3 ma postać:

$$\left\{ \begin{array}{l} \{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}, \\ \{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}, \\ \{(G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0)\}, \\ \{(G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1)\} \end{array} \right\},$$
$$\left\{ \begin{array}{l} \{(G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}, \\ \{(G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}, \\ \{(G_1,1),(G_2,1),(G_3,0),(G_4,1),(G_5,0)\}, \\ \{(G_1,1),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\} \end{array} \right\}$$

Rozpatrzmy bazę D_1 . Zgodnie z twierdzeniem 7.3-5 przestrzeń horyzontalna dla D_1 jest wypukła. Oznacza to, że możemy ją reprezentować za pomocą płotu $\langle L, R \rangle$. Do prawego krańca należą wszystkie maksymalne zbiory przedmiotów należące do elementów bazy D_1 , czyli po prostu jej transakcje. Przestrzeń zawiera wszystkie podzbiory zbiorów z krańca R , czyli L jest zbiorem pustym. Przestrzeń horyzontalna dla D_1 jest reprezentowana przez płot

$$\langle \emptyset, \{ \{(G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1)\}, \{(G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0)\}, \{(G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1)\} \} \rangle$$

Twierdzenie 7.3-6 ([27])

Niech $\{D_1, D_2\}$ będzie rodziną baz danych ze względu na decyzję dla tablicy decyzyjnej DT .

Przestrzeń wzorców wyskakujących z D_1 do D_2 jest wypukła.

Rozpatrzmy rodzinę baz DBF_4 . Zgodnie z twierdzeniem 7.3-5 przestrzeń wzorców wyskakujących z D_1 do D_2 jest wypukła. Oznacza to, że $JEP_{DBF_4}(D_1, D_2) = [L, R]$ dla pewnych rodzin L i R . Płot tej przestrzeni został wyznaczony w przykładzie w rozdziale 7.4.6 i ma postać:

$$\begin{aligned}
< \{ & \{(G_2,1)\}, \{(G_3,0),(G_5,0)\}, \{(G_3,0),(G_4,0)\}, \\
& \{(G_1,1)\}, \{(G_4,1),(G_5,0)\} \\
\{ & \{(G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}, \\
& \{(G_1,1),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}, \\
& \{(G_1,1),(G_2,1),(G_3,0),(G_4,1),(G_5,0)\} \\
& \} >
\end{aligned}$$

Licznosci i wsparcia wzorców w przestrzeni bazy D_2 zostały zamieszczone w tabelach 7.3-7 oraz 7.3-8.

Wzorzec	$count_{D_2}(X)$	$supp_{D_2}(X)$
$\{(G_2,1)\}$	4	100%
$\{(G_3,0),(G_5,0)\}$	4	100%
$\{(G_3,0),(G_4,0)\}$	3	75%
$\{(G_1,1)\}$	2	50%
$\{(G_4,1),(G_5,0)\}$	1	25%

Tab. 7.3-7. Wzorce w lewym krańcu płotu przestrzeni wzorców wyskakujących

Wzorzec	$count_{D_2}(X)$	$supp_{D_2}(X)$
$\{(G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}$	2	50%
$\{(G_1,1),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}$	1	25%
$\{(G_1,1),(G_2,1),(G_3,0),(G_4,1),(G_5,0)\}$	1	25%

Tab. 7.3-8. Wzorce w prawym krańcu płotu przestrzeni wzorców wyskakujących

7.4. Wyznaczanie wzorców wyskakujących

7.4.1. Wprowadzenie

W rozdziale tym zajmować się będziemy tematyką wyznaczania wzorców wyłaniających, przy czym skupimy się na problemie znajdowania przestrzeni wzorców wyskakujących.

Poniżej poruszam wyłącznie metody dokładne wyznaczania wzorców. W pierwszej kolejności prezentuję złożoność problemu wyznaczania wzorców wyłaniających oraz elementarne podejścia do jego rozwiązania. Następnie skupiam się na problemie wyznaczania przestrzeni wzorców wyskakujących i podaję dla niego efektywny algorytm zaproponowany w pracy [27].

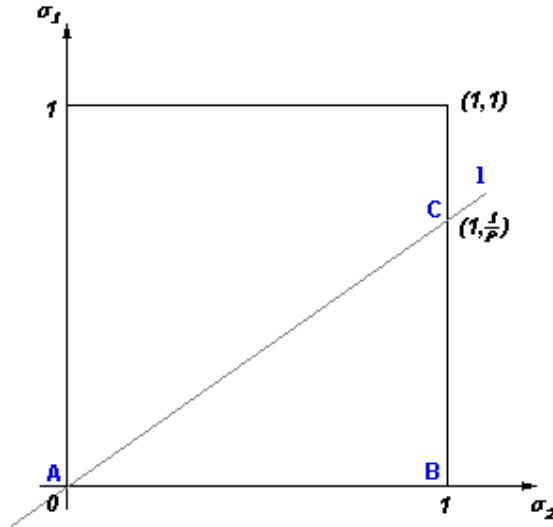
W dalszych rozważaniach zakładam, że dana jest rodzina baz danych $\{D_1, D_2\}$ o przestrzeni przedmiotów S . Zauważmy na wstępie, że przestrzeń wszystkich możliwych podzbiorów S wynosi 2^S i ma ona rozmiar $2^{|S|}$.

W prezentowanych algorytmach pomijamy problem wielkości bazy danych. Pojedyncze przejście bazy danych, najczęściej przy obliczaniu wsparcia dla danego zbioru przedmiotów, traktujemy jako operację jednostkową. Dla określenia złożoności danej metody istotne jest oszacowanie liczby takich operacji niezbędnych do otrzymania wyniku.

7.4.2. Przestrzeń wsparcia wzorców

Rozważmy problem wyznaczania przestrzeni ρ -wzorców wyłaniających z bazy D_1 do D_2 . Naszym zadaniem jest znalezienie pewnej rodziny zbiorów przedmiotów zawartej w 2^S . Zadany problem możemy w przejrzysty sposób przedstawić za pomocą płaszczyzny wsparcia.

Rozważmy kwadratowy obszar przestrzeni dwuwymiarowej opisany przez $0 \leq \sigma_1, \sigma_2 \leq 1$. Punkt obszaru (σ_1, σ_2) reprezentuje rodzinę zbiorów $\{X \in S : \text{supp}_1(X) = \sigma_1 \wedge \text{supp}_2(X) = \sigma_2\}$. Dla danego minimalnego współczynnika wsparcia ρ możemy wykreślić prostą $l : \sigma_1 = \frac{1}{\rho} \sigma_2$, która przecina prostą $\sigma_2 = 1$ w punkcie $\left(1, \frac{1}{\rho}\right)$. Wprowadźmy oznaczenia punktów $A=(0,0)$, $B=(1,0)$, $C=(1, 1/\rho)$. Suma rodzin podzbiorów odpowiadającym punktom trójkąta ABC z wyłączeniem punktu A jest szukaną przestrzenią ρ -wzorców wyłaniających z bazy D_1 do D_2 .



Rys. 5.4.3-1. Płaszczyzna wsparcia dla rodziny baz $\{D_1, D_2\}$

Zaproponowane przedstawienie problemu jest przydatne przy rozważaniu różnych typów wzorców i algorytmów z nimi związanych.

Zauważmy, że dla bazy D_1 suma rodzin podzbiorów odpowiadającym punktom odcinka osi $0 < \sigma_1 \leq 1$ jest przestrzenią horyzontalną, a punkt $\sigma_1=1$ reprezentuje zbiory o wsparciu zerowym, analogicznie dla bazy D_2 . Zgodnie z tą obserwacją zbiór wzorców wyskakujących z D_1 do D_2 jest reprezentowany przez odcinek $0 < \sigma_2 \leq 1 \wedge \sigma_1 = 0$. Punkty tego odcinka odpowiadają tym rodzinom zbiorów przedmiotów, dla których współczynnik wzrostu wsparcia z D_1 do D_2 jest nieskończony.

7.4.3. Podstawowe metody wyznaczania wzorców wyłaniających

Rozważmy problem wyznaczania przestrzeni ρ -wzorców wyłaniających z bazy D_1 do D_2 . Najprostsze rozwiązanie polega na wyczerpującym przejrzeniu wszystkich możliwych podzbiorów S i obliczeniu ich wsparć w każdej z baz. Rozpatrywany podzbiór jest wzorcem wyłaniającym, jeśli jego współczynnik wzrostu wsparcia nie spada poniżej ρ . Ponieważ przestrzeń ta zawiera $2^{|S|}$ elementów, więc systematyczne przejście możliwe jest jedynie dla małych $|S|$.

Zauważmy, że wszystkie podzbiory przedmiotów, które nie występują w żadnej transakcji danej bazy mają wsparcie zerowe w tej bazie. Obserwacja ta nasuwa prostą optymalizację, która ogranicza przestrzeń poszukiwań do przestrzeni podzbiorów wszystkich transakcji obu

baz równej $\bigcup_{T \in D_1 \cup D_2} 2^T$. Jeśli każdy element tej ograniczonej przestrzeni zostanie przejrany dokładnie raz, to złożoność zmodyfikowanego algorytmu jest niegorsza od złożoności algorytmu oryginalnego i zależy od postaci baz. Dla większych baz nie jest możliwe efektywne przechowywanie słownika dla rodziny $D_1 \cup D_2$. Ponadto zastosowana optymalizacja nie redukuje wykładniczego charakteru algorytmu i nie rozszerza wydawnie jego pola zastosowań. Nie zmienia to faktu, że podejście wyczerpujące bywa stosowane dla małych tablic decyzyjnych oraz bywa elementem zaawansowanych metod wyznaczania wzorców umożliwiając rozwiązanie podproblemu dla określonej podprzestrzeni rodziny wszystkich podzbiorów S .

Pojęcie wzorca wyłaniającego jest silnie związane z pojęciem zbioru częstego. Znacznym ulepszeniem poprzednich strategii jest oddzielne rozważenie wszystkich zbiorów częstych i rzadkich w każdej z baz. Zbiory częste można wyznaczyć między innymi za pomocą algorytmów typu Apriori. Znalezione w ten sposób rodziny podzbiorów przedmiotów o wsparciach w bazach D_1 i D_2 zawartych w dobranych odpowiednio przedziałach są szukanymi ρ -wzorcami wyłaniającymi. Należy tu zwrócić uwagę na dwa fakty. Po pierwsze rodziny zbiorów częstych przy określonym minimalnym wsparciu mogą mieć bardzo duże rozmiary, szczególnie jeśli próg jest niski. Po drugie wciąż pozostaje podprzestrzeń zbiorów rzadkich w obu bazach, która trzeba przejrzeć w sposób wyczerpujący.

Efektywne rozwiązanie pierwszego problemu dostarcza praca [13], która wykorzystuje koncepcję reprezentacji płotowej do operowania na rodzinach zbiorów przedmiotów. Autorzy systematyzują wyżej zaproponowaną strategię dekompozycji całej przestrzeni zbiorów przedmiotów. W celu uniknięcia enumeracji kolejnych zbiorów kandydujących używają algorytmu Max-Miner do znajdowania krańców rodzin zbiorów częstych oraz wykorzystują działanie różnicy płotów o specjalnej postaci.

Dalej skupimy się na wyznaczaniu przestrzeni wzorców wyskakujących. Zauważmy, że dysponując przestrzenią σ -wzorców wyłaniających dla pewnego σ możemy wyznaczyć przestrzeń wzorców wyskakujących wybierając wyłącznie te wzorce, które mają współczynnik wzrostu wsparcia równy 1. Dla rozważanego problemu można również przeformułować podane na początku podejście wyczerpujące.

Podstawowym problemem pozostaje tutaj wykładniczy rozmiar przestrzeni podzbiorów przedmiotów i potencjalnie wykładniczy rozmiar rodziny wynikowej. W rozdziale tym omawiany jest dokładnie algorytm zaproponowany w pracy [27]. Wykorzystuje on własność wypukłości przestrzeni wzorców wyskakujących oraz reprezentację płotową rodzin zbiorów.

W tej pracy nie będziemy się zajmować wyznaczaniem rodzin wzorców silnych.

7.4.4. Wyznaczanie płotów horyzontalnych

Istotnym elementem wyznaczania przestrzeni wzorców wyskakujących jest znalezienie przestrzeni horyzontalnej dla zadanej bazy danych.

Niech dana będzie baza danych $D = \{T_1, \dots, T_m\}$, przy czym $m = |D|$.

Zauważmy, że przestrzeń horyzontalna jest wypukła, gdyż każdy podzbiór jej elementu X ma wsparcie nie mniejsze niż wsparcie X . Z twierdzeń 7.3-4 i 7.3-5 wynika, że dla dowolnej przestrzeni horyzontalnej istnieje dokładnie jeden reprezentujący ją płot.

Ponieważ rodzina zawiera wszystkie zbiory o niezerowym wsparciu, więc lewy kraniec nie może wykluczać z rodziny żadnego zbioru elementów i równa się $\{\emptyset\}$. Z tego względu przestrzeń horyzontalną opisuje jednoznacznie płot postaci $\langle\{\emptyset\}, R\rangle$ dla pewnego R .

Do rozpatrywanej przestrzeni należy każdy podzbiór dowolnej transakcji bazy. Stąd prawy kraniec R wyznaczają wszystkie transakcje tej bazy danych. Ponieważ prawy kraniec jest antylańcuchem, dlatego zawiera dokładnie zbiór elementów maksymalnych dla relacji inkluzji w zbiorze wszystkich transakcji bazy D .

Poniższy algorytm buduje prawy kraniec płotu iterując kolejne transakcje bazy danych. W każdym kroku pętli ze zbioru R są usuwane elementy, które są podziorami wstawianej transakcji. W ten sposób na początku pętli zbiór R zawsze jest antylańcuchem.

```
funkcja horizon-miner(D)
R ← {T1}
dla każdego i od 2 do m
    jeśli nie istnieje takie T ∈ R, że Ti ⊂ T to
        usuń z R elementy T, takie że T ⊂ Ti
        R ← R ∪ {Ti}
zwróć  $\langle\{\emptyset\}, R\rangle$ 
```

7.4.5. Wyznaczanie różnicy płotów

Podstawową operacją wykorzystywaną przy wyznaczaniu przestrzeni wzorców wyskakujących jest różnica dwóch płotów o określonej postaci. Dotyczy ona ogólnie rodzin zbiorów, dlatego zostanie omówiona w oderwaniu od konkretnej bazy danych.

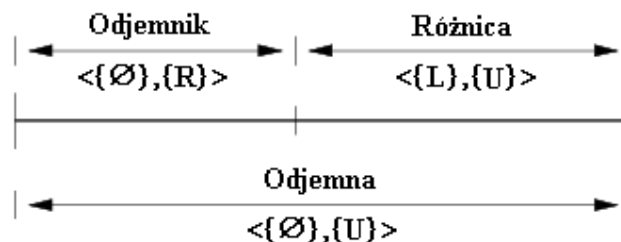
Niech dana będzie przestrzeń S .

Argumenty działania mają następującą postać:

- odjemna: $\langle\{\emptyset\}, \{U\}\rangle$
- odjemnik $\langle\{\emptyset\}, R\rangle$, gdzie $R_I = \{S_1, \dots, S_k\}$

Zachodzi $U, S_1, \dots, S_k \subset S$.

Funkcja wyznacza płot $\langle L, \{U\}\rangle$, który odpowiada przedziałowi zbiorów równemu różnicy przedziału reprezentowanego przez odjemną i rodziny reprezentowanej przez odjemnik. Zadanie algorytmu sprowadza się do wyznaczenia krańca L .



Rys. 5.4.5-1. Schemat ideowy różnicy płotów

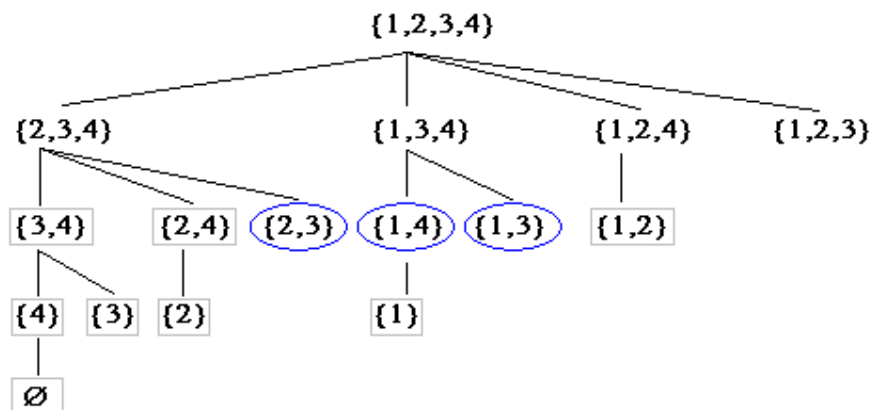
Przykład

Niech dana będzie przestrzeń liczb $U = \{1, 2, 3, 4\}$ z relacją porządku $<$ dla liczb naturalnych. Rozważmy różnicę płotu $\langle\{\emptyset\}, \{1, 2, 3, 4\}\rangle$ i płotu $\langle\{\emptyset\}, \{\{3, 4\}, \{2, 4\}, \{1, 2\}\}\rangle$.

Wynikiem odejmowania jest płot $\langle L, \{\{1,2,3,4\}\} \rangle$ reprezentujący rodzinę:
 $\{\{L\}, \{\{1,2,3,4\}\}\} = \{\{\emptyset\}, \{\{1,2,3,4\}\}\} - \{\{\emptyset\}, \{\{3,4\}, \{2,4\}, \{1,2\}\}\}.$

Ze względu na szczególną postać argumentów działanie modyfikuje wyłącznie lewy kraniec odjemnej. W celu przejrzystego przedstawienia operacji wykorzystamy drzewo enumeracji zbiorów dla rodziny 2^S .

Drzewa enumeracji są strukturami, które pozwalają na prezentację zbioru potęgowego pewnego rozważanego zbioru. Porządek węzłów i gałęzi odzwierciedla wprowadzoną w zbiorze potęgowym relację porządku. Drzewa tego typu są często stosowane do pokazania sposobu, w jaki dany algorytm rozważa kolejne elementy zbioru potęgowego. Struktury te szerzej zostały omówione w pracy [42].



Rys. 5.4.5-2. Schemat różnicy płotów $\langle \{\emptyset\}, \{1,2,3,4\} \rangle$ i $\langle \{\emptyset\}, \{\{3,4\}, \{2,4\}, \{1,2\}\} \rangle$ przedstawiony na drzewie enumeracji zbiorów

Za pomocą szarych prostokątów zostały wyróżnione zbiory odejmowane, a za pomocą niebieskich elips zbiory stanowiące lewy kraniec szukanej różnicy. Ostatecznie szukana różnica wynosi $\langle \{\{2,3\}, \{1,4\}, \{1,3\}\}, \{\{1,2,3,4\}\} \rangle$.

Schemat algorytmu ma postać:

```
funkcja border-diff ( $\langle \{\emptyset\}, \{U\} \rangle, \langle \{\emptyset\}, \{S_1, \dots, S_k\} \rangle$ )
L ←  $\{\{x\} : x = U - S_1\}$ 
dla każdego i od 2 do k
    L ←  $\{X \cup \{x\} : X \in L, x = U - S_i\}$ 
    usuń z L elementy, które nie są minimalne
zwróć  $\langle L, \{U\} \rangle$ 
```

Dokładny opis i uzasadnienie poprawności algorytmu znaleźć można w pracy [27]. Metoda zawdzięcza swoją efektywność ograniczeniu się do operacji na krańcach płotów, zamiast na elementach odpowiadających im przedziałom zbiorów. Warto dodatkowo zwrócić uwagę na fakt, że w każdej iteracji z rodziny L usuwane są wszystkie elementy nieminimalne. Etap ten wydawnie redukuje rozmiary kolejno otrzymywanych rodzin tymczasowych.

7.4.6. Wyznaczanie przestrzeni wzorców wyskakujących

Prezentowany algorytm wyznaczania przestrzeni wzorców wyskakujących omija problem potencjalnie wykładniczego rozmiaru wynikowej przestrzeni. Punktem wyjścia jest tutaj

twierdzenie 7.3-6 mówiące, że przestrzeń wzorców wyskakujących jest wypukła. Algorytm wykorzystuje reprezentację płotową do opisu przedziałów zbiorów. Istotnie, na żadnym etapie podejście nie enumeruje zbiorów danego przedziału, lecz wszelkie operacje przeprowadzane są na krańcach odpowiednich płotów.

Niech dana będzie rodzina baz danych $\{D_1, D_2\}$ o przestrzeni przedmiotów S . Rozważmy problem wyznaczania przestrzeni wzorców wyskakujących z bazy D_1 do D_2 . Zgodnie z twierdzeniem 7.3-2 szukana przestrzeń zawiera wszystkie podzbiory S , które mają niezerowe wsparcie w D_2 i zerowe wsparcie w D_1 . Idea działania algorytmu sprowadza się do wyznaczenia różnicy pomiędzy płotem przestrzeni horyzontalnej dla D_2 o postaci $\langle \{\emptyset\}, R_2 \rangle$ i płotem przestrzeni horyzontalnej dla D_1 o postaci $\langle \{\emptyset\}, R_1 \rangle$.

Wprowadźmy oznaczenia:

- $R_2 = \{A_1, \dots, A_g\}$, przy czym $g = |R_2|$
- $R_1 = \{B_1, \dots, B_h\}$, przy czym $h = |R_1|$

Ważnym elementem algorytmu jest funkcja border-diff. Narzuca ona jednak ograniczenia co do postaci dopuszczalnych argumentów, które nie pozwalają zastosować jej bezpośrednio do płotów wejściowych. Podczas, gdy odjemnik jest płotem przestrzeni horyzontalnej dla pewnej bazy danych, odjemna musi być płotem o jednoelementowym prawym krańcu.

Schemat algorytmu [27]:

```
funkcja jep-producer ( $\langle \{\emptyset\}, \{A_1, \dots, A_g\} \rangle, \langle \{\emptyset\}, \{B_1, \dots, B_h\} \rangle$ )
L  $\leftarrow \emptyset$ 
R  $\leftarrow \emptyset$ 
dla każdego j od 1 do g
    jeśli istnieje takie i, że  $A_j \subset B_i$  to
        kontynuuj
     $\langle L', R' \rangle \leftarrow \text{border-diff} (\langle \{\emptyset\}, \{A_j\} \rangle, \langle \{\emptyset\}, \{B_1, \dots, B_h\} \rangle)$ 
    L  $\leftarrow L \cup L'$ 
    R  $\leftarrow R \cup R'$ 
zwróć  $\langle L, R \rangle$ 
```

Dowód poprawności algorytmu sprowadza się do udowodnienia twierdzenia mówiącego, że suma dwóch przestrzeni wzorców wyskakujących reprezentowanych odpowiednio przez $\langle L_1, \{A_1\} \rangle$ oraz $\langle L_2, \{A_2\} \rangle$ jest przestrzenią wyskakującą reprezentowaną przez $\langle L_1 \cup L_2, \{A_1, A_2\} \rangle$. Uzasadnienie prawdziwości twierdzenia i dokładny opis działania algorytmu znajduje się w pracy [27].

Przykład

Rozpatrzmy rodzinę baz danych ze względu na decyzję $DBF_4 = \{D_1, D_2\}$ dla tablicy decyzyjnej DT_3 . Zajmiemy się wyznaczeniem przestrzeni wzorców wyskakujących $JEP_{DBF_4}(D_1, D_2)$. Rodzina DBF_4 ma postać:

$$\{ \{ \{ (G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1) \}, \\ \{ (G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1) \}, \\ \{ (G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0) \}, \\ \{ (G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1) \} \} , \\ \{ \{ (G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \}, \\ \{ (G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \}, \\ \{ (G_1,1),(G_2,1),(G_3,0),(G_4,1),(G_5,0) \}, \\ \{ (G_1,1),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \} \} \}$$

Prawe krańce przestrzeni horyzontalnych, które są argumentami funkcji jep-producer mają postać:

$$R_1 = \{ \{ (G_1,0),(G_2,0),(G_3,0),(G_4,1),(G_5,1) \}, \\ \{ (G_1,0),(G_2,0),(G_3,1),(G_4,0),(G_5,0) \}, \\ \{ (G_1,0),(G_2,0),(G_3,1),(G_4,1),(G_5,1) \} \}$$

$$R_2 = \{ \{ (G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \}, \\ \{ (G_1,1),(G_2,1),(G_3,0),(G_4,1),(G_5,0) \}, \\ \{ (G_1,1),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \} \}$$

Przyjrzymy się pierwszej iteracji algorytmu dla $j=1$.

W tym przypadku mamy $A_j = \{(G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}$.

Nie istnieje taki element zbioru R_1 , który zawierałby zbiór A_j , więc wykonane zostanie działanie różnicy: $[\{\emptyset\}, \{(G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0)\}] - [\{\emptyset\}, R_1]$. Wynikiem różnicy jest przedział reprezentowany przez płot $\langle L', R' \rangle$:

$$\langle \{ \{ (G_2,1) \}, \\ \{ (G_3,0),(G_5,0) \}, \\ \{ (G_3,0),(G_4,0) \} \\ \{ \{ (G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \} \} \} \rangle$$

Przedział ten inicjalizuje również pusty dotąd płot $\langle L, R \rangle$. Zauważmy, że oba krańce zbliżają się do postaci wynikowej, która otrzymana zostanie po 3 iteracjach. Ostatecznie algorytm wyznacza płot przestrzeni wzorców wyskakujących równy:

$$\langle \{ \{ (G_2,1) \}, \{ (G_3,0),(G_5,0) \}, \{ (G_3,0),(G_4,0) \}, \\ \{ (G_1,1) \}, \{ (G_4,1),(G_5,0) \} \\ \{ \{ (G_1,0),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \}, \\ \{ (G_1,1),(G_2,1),(G_3,0),(G_4,0),(G_5,0) \}, \\ \{ (G_1,1),(G_2,1),(G_3,0),(G_4,1),(G_5,0) \} \} \} \rangle$$

Algorytm zwraca w wyniku płot $\langle L, R \rangle$ odpowiadający przestrzeni wszystkich wzorców wyskakujących z D_1 do D_2 . Otrzymana przestrzeń może mieć potencjalnie rozmiar wykładniczy. Zastanówmy się nad przydatnością wyniku. Po pierwsze analizując krańce płotu można oszacować wielkość otrzymanej przestrzeni. W przypadku małych przestrzeni można wygenerować wszystkie wzorce wyskakujące przechodząc gałęzie drzewa enumeracji zbiorów. W przeciwnym przypadku rezultat może być wykorzystany pośrednio.

Przed wszystkim dużo informacji zawartych jest w zbiorach należących do krańców płotu. Zauważmy, że zbiory z lewego krańca przedziału są wzorcami wyskakującymi o największym wsparciu w D_2 . Z tego względu będą one prawdopodobnie często występowały w transakcjach spoza rozważanych baz i mogą być wykorzystane do budowy klasyfikatorów. Zbiory należące do prawego krańca są najbardziej specyficznymi, maksymalnymi wzorcami wyłaniającymi w przestrzeni wynikowej. Mogą one dostarczyć wiedzy na temat istnienia specyficznych zależności, których najczęściej nie można uzyskać za pomocą innych znanych metod, np. statystycznych. Ponadto, dla danego zbioru przedmiotów z przestrzeni S można sprawdzić, czy należy on do wyznaczonej przestrzeni wzorców za pomocą co najwyżej $|L|+|R|$ sprawdzeń zachodzenia inkluzji dwóch zbiorów.

7.4.7. Analiza efektywności

W przypadku zadań odkrywania wiedzy rozpatrywane dane charakteryzują się zwykle wysoką wymiarowością. Z tego względu przestrzeń rozpatrywanych przedmiotów jest często duża, a co za tym idzie jej zbiór potęgowy nie może być systematycznie przejrany metodami wyczerpującymi.

W rozdziale tym omówimy złożoność funkcji wykorzystywanych do wyznaczenia przestrzeni wzorców wyskakujących. Na wstępie należy zaznaczyć, że autorzy algorytmów nie podają formalnej analizy złożoności, lecz prezentują wyniki eksperymentów przeprowadzonych dla różnych zbiorów danych [27]. Ograniczymy się do podsumowania tych rezultatów.

Niech dana będzie rodzina baz danych $\{D_1, D_2\}$ o przestrzeni przedmiotów S . Rozważmy problem znajdowania wzorców wyskakujących z D_1 do D_2 .

Funkcja horizon-miner wywoływana jest jednokrotnie dla każdej z baz danych. Rozpatruje ona kolejno transakcje bazy danych. Dla każdej transakcji utrzymywany jest niezmiennik, że rodzina tymczasowa zbiorów jest antyłańcuchem, co realizowane jest przez przejrzenie tej rodziny i usunięcie odpowiednich zbiorów. Krok ten jest liniowy ze względu na chwilowy rozmiar rodziny tymczasowej. Przyjeliśmy tu za operację jednostkową sprawdzenie inkluzji pomiędzy dwoma zbiorami.

Podsumowując, algorytm działa w pesymistycznym czasie $O(|D|^2)$. Przypadek pesymistyczny zachodzi, gdy każda transakcja występuje tylko raz w D i gdy zbiór transakcji z D tworzy antyłańcuch.

Algorytm jep-producer składa się z jednej głównej pętli, która przeprowadza analogiczne operacje dla każdego elementu prawego krańca płotu przestrzeni horyzontalnej D_2 , w szczególności operacje wyznaczania różnicy płotów. Ponieważ utrzymywanie rodzin tymczasowych ma poboczne znaczenie dla złożoności, więc decydujący wpływ na efektywność algorytmu ma funkcja border-diff.

Czas potrzebny na wyznaczenie różnicy płotów $\langle \{\emptyset\}, \{U\} \rangle, \langle \{\emptyset\}, R \rangle$ za pomocą funkcji border-diff zależy od zbioru U i rodziny R . Badanie efektywności czasowej zostało przeprowadzone dla różnych zbiorów danych w pracy [27]. Eksperyment sprowadzał się do porównania czasu wykonania procedury border-diff z czasem wykonania algorytmu naiwnego. Algorytm naiwny przeprowadzał odejmowanie poprzez bezpośrednie wyznaczenie wszystkich elementów przedziałów reprezentowanych przez oba płoty. Eksperyment został przeprowadzony na komputerze PC z procesorem Pentium III i 512MB pamięci RAM. Czas działania algorytmu naiwnego został oszacowany przez funkcję $0,1 * 2^{|U|} / 2^9$ opartej na fakcie, że wykonanie tego algorytmu dla $|U|=9$ trwało $0,1s$. Zbiory danych pochodzą z UCI Repository [58].

Zbiór danych	$ U $	$ R $	$ L $	Czas procesora dla algorytmu border-diff	Czas procesora dla algorytmu naiwnego
tic-tac-toe	9	332	4	0,01s	0,1s
Mushroom	22	3916	86	0,65s	>5ms
Chess	36	1527	42	0,15s	>30ms
Sonar	60	111	418	1,82s	>2h

Tab. 7.4.7-1. Porównanie efektywności czasowej algorytmu border-diff z algorytmem naiwnym dla różnych zbiorów danych

Niestety dla dużych zbiorów danych i niektórych argumentów wejściowych czas wykonania funkcji border-diff może okazać się za długi. Dla tych przypadków można zdecydować się na wyznaczenie jedynie części poszukiwanej przestrzeni. Propozycje rozwiązań dostępne są w pracy [27].

Ogólnie warto podkreślić, że liczba transakcji w rozpatrywanych bazach danych ma drugorzędne znaczenie na czas wyznaczania przestrzeni wzorców wyskakujących w stosunku do wymiarowości problemu. Autorzy podają, że dla ustalonej wymiarowości czas działania wzrastał liniowo wraz z przyrostem liczby transakcji w bazach.

8. Zależność reduktów i wzorców wyskakujących

8.1. Wprowadzenie

W rozdziale tym zajmiemy się wzajemnym porównaniem zbioru reduktów danej tablicy decyzyjnej z przestrzenią wzorców wyskakujących dla rodziny baz danych wyznaczonej dla tej tablicy decyzyjnej. Przedstawimy przesłanki teoretyczne oraz zaproponujemy metody analizy wyników eksperymentalnych. Omówienie przeprowadzonych eksperymentów i podsumowanie rezultatów znajdują się w rozdziale 9.

8.2. Rozważania formalne

Zarówno pojęcie reduktu, jak i wzorca wyskakującego związane jest z pewnym zbiorem atrybutów, który charakteryzuje się własnością dobrego rozróżniania obiektów według klasy decyzyjnej. Pomimo tej zbieżności, zależność obu pojęć nie została dotąd formalnie wyprowadzona. W pracy ograniczymy się do analizy tablic deterministycznych.

Definicja 8.2-1

Tablica decyzyjna $DT=(U,A,C,\{d\})$ jest:

- **deterministyczna**, jeśli $\forall u_1, u_2 \in U : (u_1 = u_2) \Rightarrow d(u_1) = d(u_2)$
- **niedeterministyczna**, w przeciwnym przypadku.

Rozważania można rozszerzyć na przypadek dowolnej tablicy niedeterministycznej. W tym celu należy usunąć wszystkie obiekty, które wprowadzają niedeterminizm. Dla danej tablicy $DT=(U,A,C,\{d\})$ należy wyznaczyć tablicę decyzyjną $DT'=(U',A,C,\{d\})$, przy czym $U' = \{u \in U : \neg \exists (v \in U)((u = v) \wedge d(u) \neq d(v))\}$.

W rozdziale skupimy się na podstawowych obserwacjach dotyczących wyróżniania jednej wybranej klasy spośród reszty. Rozważmy deterministyczną tablicę decyzyjną $DT=(U,A,C,\{d\})$, dla której $V_d=\{d_1, \dots, d_p\}$. Bez straty ogólności wybierzmy klasę d_1 jako klasę odniesienia. W ten sposób otrzymamy tablicę decyzyjną, w której wyróżniamy obiekty z dwóch klas: należące do d_1 i należące do jednej z pozostałych klas DT . Możemy zdefiniować

tablicę $DT'=(U,A,C,\{d'\})$, taką że $d':U \rightarrow \{1,2\}$ i $d'(u) = \begin{cases} 1 & d(u) = d_1 \\ 2 & d(u) \neq d_1 \end{cases}$.

Dla późniejszej wygody wprowadźmy oznaczenie zbiorów obiektów oraz systemów informacyjnych należących odpowiednio do każdej z klas. Dla $i=1,2$ mamy $U_i = \{u \in U : d'(u) = i\} = (d')^{-1}(i)$ oraz $IS_i = (U_i, C)$.

Definicja 8.2-2

Niech:

- $IS=(U,A)$ będzie systemem informacyjnym
- $u \in U$
- $X \subset A$

Zbiór przedmiotów generowany przez obiekt względem zbioru atrybutów w systemie informacyjnym oznacza zbiór $\{(a, a(u))\}_{a \in X}$.

Zbiór przedmiotów generowany przez obiekt względem zbioru atrybutów w systemie informacyjnym będziemy oznaczać przez $I_{IS}(u, X)$.

Rozpatrzmy rodzinę baz danych względem decyzji $DBF' = \{D_1, D_2\}$ dla tablicy DT' , przy czym baza D_1 odpowiada klasie 1, a D_2 – klasie 2. Naszym zadaniem jest znalezienie powiązania pomiędzy zbiorem reduktów $RED_{DT'}(C)$ i zbiorem $JEP_{DBF'}(D_1, D_2)$. Związek pomiędzy reduktami i pewnym zbiorem wzorców wyskakujących wyraża następujące twierdzenie.

Twierdzenie 8.2-1

Niech dana będzie deterministyczna tablica decyzyjna DT' i rodzina baz danych DBF' dla tej tablicy zdefiniowane jak powyżej, redukt $R \in RED_{DT'}(C)$ oraz obiekt $u_2 \in U_2$.

Zbiór $I_{IS_2}(u_2, R)$ jest wzorcem wyskakującym z D_1 do D_2 .

Dowód:

Rozpatrzmy dwa systemy informacyjne $IS_1 = (U_1, C)$ i $IS_2 = (U_2, C)$. Systemy IS_1 i IS_2 wyznaczają odpowiednio bazy danych D_1 i D_2 . Zgodnie z definicją 8.2-2 wprowadźmy oznaczenie zbioru $X = I_{IS_2}(u_2, R) = \{(a, a(u_2))\}_{a \in R}$.

Ponieważ $u_2 \in U_2$, to odpowiadająca mu transakcja $T_{IS_2}(u_2) \in D_2$. Ponieważ $X \subset T_{IS_2}(u_2)$, więc $supp_{D_2}(X) > 0$.

Rozpatrzmy teraz dowolny obiekt $u_1 \in U_1$. Tablica DT' jest deterministyczna, stąd $u_1 \neq u_2$. Redukt R jest zbiorem rozróżniającym dla zbioru atrybutów C . Ponieważ $u_1 \neq u_2$ i $d(u_1) \neq d(u_2)$, więc $(u_1, u_2) \notin IND_{DT'}(R)$ i $(u_2, u_1) \notin IND_{DT'}(R)$. Oznacza to, że istnieje takie $a \in R$, że $a(u_1) \neq a(u_2)$. Innymi słowy dwa obiekty z innych klas muszą być rozróżnialne za pomocą reduktu, tzn. muszą mieć różne wartości dla przynajmniej jednego atrybutu należącego do tego reduktu. Wynika stąd, że $X \not\subset T_{IS_1}(u_1)$.

Ponieważ obiekt u_1 był wybrany w sposób dowolny, więc żadna z transakcji bazy danych D_1 nie zawiera X . Wynika stąd, że $supp_{D_1}(X) = 0$.

Biorąc pod uwagę wartości wsparcia zbioru X w obu bazach, na podstawie twierdzenia 7.3-3 wnioskujemy, że zbiór przedmiotów X jest wzorcem wyskakującym z D_1 do D_2 . □

Twierdzenie pokazuje, że dowolny redukt tablicy decyzyjnej DT' może być wykorzystany do otrzymania pewnego zbioru wzorców wyskakujących z D_1 do D_2 poprzez iterowanie kolejnych obiektów klasy 2. W ten sposób można wygenerować pewną podprzestrzeń przestrzeni $JEP_{DBF'}(D_1, D_2)$. Fakt ten precyzuje następne twierdzenie.

Twierdzenie 8.2-2

Niech dana będzie deterministyczna tablica decyzyjna DT' i rodzina baz danych DBF' dla tej tablicy zdefiniowane jak powyżej.

Zachodzi inkluzja $\{I_{IS_2}(u, R) : u \in U_2 \wedge R \in RED_{DT'}(C)\} \subset JEP_{DBF'}(D_1, D_2)$.

Dowód:

Oznaczmy $JF = \{(a, a(u))\}_{a \in R} : u \in U_2 \wedge R \in RED_{DT'}(C)\}$.

Weźmy dowolny redukt $R \in RED_{DT'}(C)$ i obiekt $u_2 \in U_2$, które wyznaczają jednoznacznie element $X = I_{IS_2}(u_2, R) = \{(a, a(u_2))\}_{a \in R}$ rodziny JF . Z twierdzenia 8.2-1 wynika, że element

X jest wzorcem wyskakującym z D_1 do D_2 . Ze względu na dowolność wyboru elementu rodziny JF wnioskujemy, że wszystkie jej elementy należą do $JEP_{DBF}(D_1, D_2)$, co jest równoważne z tezą. □

Zauważmy, że sposób generowania podprzestrzeni wzorców JF na podstawie wszystkich obiektów klasy 2 pociąga za sobą fakt, że $\sum_{J \in JF} \text{supp}_{D_2}(J) = 1$.

Powyższe twierdzenia pozostają prawdziwe nie tylko dla reduktu, ale również dla dowolnego zbioru rozróżniającego. Podkreślimy jednocześnie, że poruszają one wyłącznie charakter otrzymywanych zbiorów przedmiotów. nierozstrzygnięta pozostaje kwestia, jak duże wsparcie ma każdy z wzorców wyskakujących w bazie D_2 . W szczególności nie wskazują one na powiązanie pomiędzy reduktami a wzorcami wyznaczającymi płot przestrzeni $JEP_{DBF}(D_1, D_2)$.

Przykład

Rozważmy tablicę decyzyjną DT_3 i rodzinę baz danych $DBF_4 = \{D_1, D_2\}$. Baza D_1 odpowiada klasie Z, a baza D_2 – klasie R.

	G_1	G_2	G_3	G_4	G_5	d
H_1	0	0	0	1	1	Z
H_2	0	0	0	1	1	Z
H_3	0	0	1	0	0	Z
H_4	0	0	1	1	1	Z
H_5	0	1	0	0	0	R
H_6	0	1	0	0	0	R
H_7	1	1	0	1	0	R
H_8	1	1	0	0	0	R

Tab. 8.2-1. Tablica decyzyjna DT_3

Zgodnie z poprzednimi rozważaniami mamy, że $RED_{DT_3}(C) = \{\{G_2\}, \{G_3, G_5\}, \{G_1, G_3, G_4\}\}$.

Zgodnie z twierdzeniem 8.2-2 dla zbioru reduktów możemy wygenerować rodzinę F wzorców wyskakujących z D_1 do D_2 . Rodzina F ma postać:

$$\{ \{ (G_2, 1) \}, \{ (G_3, 0), (G_5, 0) \}, \{ (G_1, 0), (G_3, 0), (G_4, 0) \}, \{ (G_1, 1), (G_3, 0), (G_4, 1) \}, \{ (G_1, 1), (G_3, 0), (G_4, 0) \} \}$$

Z poprzednich rozważań mamy przestrzeń $JEP(D_1, D_2)$ reprezentowaną w postaci płotu

$$< \{ \{ (G_2, 1) \}, \{ (G_3, 0), (G_5, 0) \}, \{ (G_3, 0), (G_4, 0) \}, \{ (G_1, 1) \}, \{ (G_4, 1), (G_5, 0) \} \}, \\ \{ \{ (G_1, 0), (G_2, 1), (G_3, 0), (G_4, 0), (G_5, 0) \}, \{ (G_1, 1), (G_2, 1), (G_3, 0), (G_4, 0), (G_5, 0) \}, \{ (G_1, 1), (G_2, 1), (G_3, 0), (G_4, 1), (G_5, 0) \} \} >$$

Tezę twierdzenia 8.2-2 pokażemy na wybranych elementach F , chociaż jest ona prawdziwa dla całej rodziny F .

Wzorcami wyskakującymi są między innymi:

- $\{(G_2, 1)\}$, gdyż:
 $\{(G_2, 1)\} \leq \{(G_2, 1)\} \leq \{(G_1, 0), (G_2, 1), (G_3, 0), (G_4, 0), (G_5, 0)\}$
- $\{(G_3, 0), (G_5, 0)\}$, gdyż:
 $\{(G_3, 0), (G_5, 0)\} \leq \{(G_3, 0), (G_5, 0)\} \leq \{(G_1, 1), (G_2, 1), (G_3, 0), (G_4, 0), (G_5, 0)\}$
- $\{(G_1, 1), (G_3, 0), (G_4, 1)\}$, gdyż:
 $\{(G_1, 1)\} \leq \{(G_1, 1), (G_3, 0), (G_4, 1)\} \leq \{(G_1, 1), (G_2, 1), (G_3, 0), (G_4, 1), (G_5, 0)\}$

Dla wygody zapisu wprowadźmy pojęcie wiążące zbiór przedmiotów z odpowiadającym mu zbiorem atrybutów.

Definicja 8.2-3

Niech $IS=(U, A)$ będzie systemem informacyjnym, S_{IS} przestrzenią przedmiotów i zbiór przedmiotów $X = \{(a_1, v_1), \dots, (a_k, v_k)\} \subset S_{IS}$.

Zbiór atrybutów związany ze zbiorem przedmiotów X oznacza zbiór $\{a_1, \dots, a_k\}$.

Zbiór atrybutów związany ze zbiorem przedmiotów X oznaczamy przez $attr(X)$.

Baza danych jest budowana na systemie informacyjnym, a jej transakcje odpowiadają obiektom. Z tego względu transakcja zawiera przedmioty, z których każdy związany jest z innym atrybutem. Przejdźmy do rodziny baz DBF' zbudowanej na tablicy DT' .

Twierdzenie 8.2-3

Niech dana będzie deterministyczna tablica decyzyjna $DT'=(U, A, C, \{d\})$ i rodzina baz danych DBF' dla tej tablicy, zdefiniowane jak powyżej, oraz rodzina wzorców wyskakujących $JF \subset JEP_{DBF'}(D_1, D_2)$.

Jeśli z każdym wzorcem $J \in JF$ związany jest ten sam zbiór atrybutów P i $\sum_{J \in JF} supp_{D_2}(J) = 1$,

to P jest zbiorem rozróżniającym w tablicy decyzyjnej DT' .

Dowód:

Rozważmy twierdzenie przeciwstawne. Niech $P \subset C$ nie będzie zbiorem rozróżniającym dla tablicy DT' . Z definicji 6.2-7 wynika, że $IND_{DT'}(P) \neq IND_{DT'}(C)$. Ponieważ zbiór C rozróżnia nie mniej par obiektów co zbiór P , więc istnieje taka para obiektów $(u_1, u_2) \in U \times U$, dla której: $(u_1, u_2) \in IND_{DT'}(P)$ i $(u_1, u_2) \notin IND_{DT'}(C)$. Z definicji 6.2-1 wynika:

$$\begin{aligned} & ((\forall a \in P : a(u_1) = a(u_2)) \vee d(u_1) = d(u_2)) \wedge ((\exists a \in C : a(u_1) \neq a(u_2)) \wedge d(u_1) \neq d(u_2)) \Leftrightarrow \\ & (\forall a \in P : a(u_1) = a(u_2)) \wedge (\exists a \in C : a(u_1) \neq a(u_2)) \wedge d(u_1) \neq d(u_2) \Rightarrow \\ & (\forall a \in P : a(u_1) = a(u_2)) \wedge d(u_1) \neq d(u_2) \end{aligned}$$

Bez straty ogólności załóżmy, że $d(u_1)=1$ i $d(u_2)=2$. Niech funkcje T_1 i T_2 oznaczają odpowiednio przyporządkowania transakcji dla baz D_1 i D_2 . Zachodzi $T_1(u_1) \in D_1$ i $T_2(u_2) \in D_2$. Ponieważ zachodzi $\forall a \in P : a(u_1) = a(u_2)$, więc zbiór przedmiotów $X = I(u_2, P)$ należy do transakcji $T_1(u_1)$ i $T_2(u_2)$. Wynika stąd, że $supp_{D_1}(X) > 0$ i na podstawie twierdzenia 7.3-2 zbiór X nie jest wzorcem wyskakującym z D_1 do D_2 . Oznacza to, że $X \notin JF$. Stąd, dla każdego

$J \in JF$ istnieje, takie $a \in P$, że $(a, v_a) \in J$ i $v_a \neq a(u_2)$. Oznacza to jednocześnie, że dla każdego wzorca $J \in JF$: $J \not\subset T_2(u_2)$.

Wzorce rodziny JF są związane z tym samym zbiorem przedmiotów. Stąd, dla każdego obiektu $u \in U_2$ istnieje co najwyżej jeden $J \in JF$, taki że $J \subset T_2(u)$.

Podsumowując, żadne dwa wzorce nie są wspierane przez tę samą transakcję oraz żaden wzorec rodziny nie jest wspierany przez transakcję $T_2(u)$. Wynika stąd, że $\sum_{J \in JF} \text{count}_{D_2}(J) < |D_2|$ i dalej $\sum_{J \in JF} \text{supp}_{D_2}(J) < 1$. Ostatecznie $\sum_{J \in JF} \text{supp}_{D_2}(J) \neq 1$, co dowodzi tezy twierdzenia przeciwnego. Rozumowanie jednocześnie potwierdza prawdziwość rozważanego twierdzenia prostego. □

Dodajmy, że jeśli nie istnieje rodzina JF' wzorców wyskakujących związanych z tym samym podzbiorem właściwym zbioru atrybutów P , dla której $\sum_{J \in JF'} \text{supp}_{D_2}(J) = 1$, to P jest reduktom.

Przykład

Rozważmy tablicę decyzyjną DT_3 i rodzinę baz danych $DBF_4 = \{D_1, D_2\}$. Baza D_1 odpowiada klasie Z, a baza D_2 – klasie R.

	G_1	G_2	G_3	G_4	G_5	d
H_1	0	0	0	1	1	Z
H_2	0	0	0	1	1	Z
H_3	0	0	1	0	0	Z
H_4	0	0	1	1	1	Z
H_5	0	1	0	0	0	R
H_6	0	1	0	0	0	R
H_7	1	1	0	1	0	R
H_8	1	1	0	0	0	R

Tab. 8.2-2. Tablica decyzyjna DT_3

Wsparcia wzorców wyskakujących z D_1 do D_2 rodziny $JF = \{(G_2, 1), (G_4, 0)\}, \{(G_2, 1), (G_4, 1)\}$ sumują się do 1. Zbiór $\{G_2, G_4\}$ jest rozróżniający w DT_3 .

Wzorec $J = \{(G_1, 0), (G_2, 0)\}$ jest wyskakujący z D_2 do D_1 . Dodatkowo $\text{supp}_{D_1}(J) = 1$. Zbiór atrybutów $\{G_1, G_2\}$ jest rozróżniający. Zauważmy podobnie, że wzorec $J = \{(G_2, 0)\}$ również jest wyskakujący z D_2 do D_1 i ma pełne wsparcie w D_1 , a zbiór $\{G_2\}$ jest reduktom względny tablicy DT_3 . Przykłady pokazują, że otrzymywane zbiory atrybutów nie muszą być niezależne z punktu widzenia rozróżnialności obiektów.

8.3. Prowadzenie analizy wyników

8.3.1. Wprowadzenie

Zarówno redukt, jak i wzorec wyskakujący wskazują pewne szczególne zbiory atrybutów. Obserwując zbiór reduktów i przestrzeń wzorców słuszną wydaje się intuicja, że atrybuty o istotnym znaczeniu dla rozróżniania obiektów będą wyjątkowo często występować w reduktach i schematach wzorców. Zgodnie z tą ideą możemy zdefiniować miary istotności pojedynczych atrybutów odwołujące się do częstotliwości ich występowania w zbiorach

odpowiedniej przestrzeni. Za pomocą podstawowych narzędzi statystycznych można dokonać wzajemnego porównania tych miar oraz odnieść do oceny atrybutów proponowanej w rozdziale 4.

8.3.2. Przypadek reduktów

Założmy, że dysponujemy zbiorem reduktów RF dla danej tablicy decyzyjnej DT' . Wybierając dowolny redukt możemy zredukować tablicę decyzyjną poprzez zawężenie zbioru atrybutów warunkowych. Nowy zbiór atrybutów zachowuje informacje na temat klasyfikacji obiektów. Sugeruje to, że atrybuty należące do reduktu mogą mieć nawet oddzielnie duży wpływ na wartość atrybutu decyzyjnego.

Założenie 1

Redukt dostarcza nam informacji jedynie o związku pewnego zbioru atrybutów z klasą decyzyjną. Nie wskazuje jednak na to, jak ta informacja jest rozdystrybuowana pomiędzy należące do niego atrybuty. Określając znaczenie pojedynczego atrybutu na podstawie analizy pojedynczego reduktu będziemy zakładać, że informacja klasyfikująca jest rozłożona równomiernie pomiędzy atrybuty należące do reduktu.

Każdy redukt w sposób całkowity rozróżnia obiekty tablicy. Można powiedzieć, że zawiera tę samą informację klasyfikującą. Stąd atrybuty należące do reduktów o większych rozmiarach mają mniejszy związek z klasą decyzyjną.

Założenie 2

Możemy spodziewać się, że atrybuty najistotniejsze będą występować w wielu reduktach. Informacja niesiona przez taki atrybut może być uzupełniona przez wiele zbiorów słabo związanych z hipotezą atrybutów tak, aby razem stanowiły redukt. Stąd atrybut będziemy uważać za tym istotniejszy, im częściej występuje on w reduktach tablicy decyzyjnej. Zauważmy, że w ten sposób obejmujemy informacje cząstkowe na temat znaczenia atrybutów z perspektyw różnych reduktów.

Powyższe założenia można podsumować prostą formułą oszacowującą znaczenie atrybutu:

$$attrval_{RED}(a) = \frac{1}{MAX_{RED}} attrval'_{RED}(a)$$

gdzie:

$$attrval'_{RED}(a) = \sum_{R \in RF} f_{RED}(a, R)$$

$$MAX_{RED} = \max_{a \in C} (attrval'_{RED}(a)) \quad f_{RED}(a, R) = \begin{cases} \frac{1}{|R|} & a \in R \\ 0 & a \notin R \end{cases}$$

Ze względów efektywnościowych rzadko dysponujemy pełnym zbiorem reduktów $RF = RED_{DT}(C)$. Zauważmy jednak, że zgodnie z założeniem 1 cząstkowa ocena wartości atrybutów jest mniejsza dla reduktów liczniejszych. Najistotniejsze informacje o atrybutach niosą redukt minimalne. Zgodnie z zaproponowaną formułą również największy wkład do oszacowania znaczenia atrybutu ma jego występowanie w reduktach o małej liczbie elementów.

Z powyższych względów możemy ograniczyć zbiór RF wyłącznie do reduktów o rozmiarach nieprzekraczających pewnego stałego rozmiaru. Ograniczenie będziemy dobierać w zależności od wielkości tablicy decyzyjnej DT .

Dla małych tablic decyzyjnych możemy zastosować algorytm dokładny i obliczyć współczynniki znaczenia dla wszystkich atrybutów warunkowych ze zbioru C . Najczęściej jednak musimy posługiwać się rozwiązaniami przybliżonymi.

Metoda wykorzystująca dekompozycję przestrzeni pozwala znajdować redukty o ograniczonym od góry rozmiarze. Z tego względu może być z powodzeniem stosowana do budowy zbioru RF wykorzystywanego później do szacowania znaczenia atrybutów.

Przykład

Dla tablicy decyzyjnej DT_3 wyznaczyliśmy cały zbiór reduktów za pomocą algorytmu dokładnego exact-reduct-finder. Przyjeliśmy $RF = RED_{DT_3}(C) = \{\{G_2\}, \{G_3, G_5\}, \{G_1, G_3, G_4\}\}$.

Wyliczone wartości funkcji $attrval_{RED}$ znajdują się w tabeli 8.3.2-1.

Atrybut	G_1	G_2	G_3	G_4	G_5
$attrval_{RED}$	0,333	1,000	0,833	0,333	0,500

Tab. 8.3.2-1. Wartości funkcji $attrval_{RED}$ dla atrybutów tablicy TD_3

8.3.3. Przypadek wzorców wyskakujących

Rozważmy rodzinę baz danych DBF' i pewien zbiór JF wzorców wyskakujących z D_1 do D_2 . Wzorec wyskakujący jest zbiorem przedmiotów, który występuje w transakcjach jednej bazy, a jest nieobecny w transakcjach drugiej bazy. Każdy przedmiot wzorca jest związany z określonym poziomem pewnego atrybutu. Można się spodziewać, że atrybut ten ma istotne znaczenie dla rozróżniania obiektów, co wiąże go z hipotezą docelową. Zdefiniowanie miary istotności atrybutów za pomocą analizy zbioru wzorców wyskakujących poprzedzimy sformułowaniem odpowiednich założeń. Są one analogiczne do założeń poczynionych dla reduktów.

Założenie 1

Każdy wzorec niesie ze sobą ilość informacji wprost proporcjonalną do liczności jego występowania D_2 , gdyż wskazuje ona na liczbę zawierających go transakcji z D_2 , które pozwala odróżnić od wszystkich innych transakcji z bazy D_1 . Przenosząc to na grunt tablicy decyzyjnej DT' pozwala on na rozróżnienie odpowiadających transakcjom obiektów klasy 2 od wszystkich innych klasy 1.

Założenie 2

Obserwując pojedynczy wzorec wyskakujący J nie jesteśmy w stanie określić, które z należących do niego przedmiotów są najbardziej charakterystyczne dla klasy 2. Stąd, nie możemy łatwo wnioskować, jakie atrybuty z $attr(J)$ są najistotniejsze dla rozróżniania obiektów. Nie mając informacji na temat siły wzorca, możemy stwierdzić, że każdy niesie ze sobą tę samą część informacji pozwalającej rozróżnić niektóre obiekty klasy 2 od wszystkich klasy 1. Określając znaczenie pojedynczego atrybutu na podstawie analizy pojedynczego wzorca wyskakującego J będziemy zakładać, że informacja klasyfikująca jest rozłożona równomiernie pomiędzy atrybuty $attr(J)$. Stąd będziemy przyjmować, że atrybuty związane z wzorcami o większych rozmiarach mają mniejszy wpływ na klasę decyzyjną.

Założenie 3

Możemy spodziewać się, że najbardziej charakterystyczne dla klasy 2 przedmioty będą obecne w wielu wzorcach wyłaniających. Występowanie przedmiotu pociąga za sobą występowanie związanego z nim atrybutu. Przy czym ten sam atrybut może być związany z różnymi przedmiotami.

Będziemy uważać, że atrybut jest tym istotniejszy im częściej występuje on w zbiorach atrybutów wzorców należących do JF . Obejmujemy w ten sposób informacje cząstkowe na temat znaczenia atrybutów z perspektywy różnych wzorców wyskakujących.

Założenia te realizuje następująca funkcja:

$$attrval_{JEP}(a) = \frac{1}{MAX_{JEP}} attrval_{JEP}'(a)$$

gdzie:

$$attrval_{JEP}'(a) = \sum_{J \in JF} f_{JEP}(a, J)$$

$$MAX_{JEP} = \max_{a \in C} (attrval_{JEP}'(a)) \quad f_{JEP}(a, J) = \begin{cases} \frac{count_{D_2}(J)}{|attr(J)|} & a \in attr(J) \\ 0 & a \notin attr(J) \end{cases}$$

Obliczenie miary dla atrybutów tablicy DT' wymaga przeanalizowania wzorców z JF . Jeśli $|JF|$ ma małe rozmiary, to może być zadana explicite i wtedy funkcję $attrval_{JEP}$ można wyznaczyć przez przejście elementów JF .

Jednak potencjalnie rodzina JF może mieć rozmiar wykładniczy. W tym wypadku musimy ograniczyć się do analizy rodzin wypukłych, które można opisać za pomocą reprezentacji płotowej lub wybranych zbiorów wzorców.

W naszych rozważaniach będziemy przyjmować, że dla danej rodziny DBF' jesteśmy w stanie wyznaczyć przestrzeń wzorców wyskakujących w reprezentacji płotowej. Za najistotniejsze wzorce wyskakujące uznaje się wzorce należące do lewego krańca płotu przestrzeni. Z tego względu przy obliczeniu funkcji oceny będziemy rozważać właśnie tę rodzinę. Podsumowując, mamy: $JEP_{DBF'}(D_1, D_2) = [L, R]$ i $JF = L$.

Przykład

Dla tablicy decyzyjnej DBF_4 wyznaczyliśmy całą przestrzeń wzorców wyskakujących za pomocą algorytmu jep-producer. Przyjęliśmy $JF = JEP_{DBF_4}(D_1, D_2)$. Płot $\langle L, R \rangle$ tej przestrzeni ma postać:

$$\begin{aligned} < \{ \{ (G_2, 1) \}, \{ (G_3, 0), (G_5, 0) \}, \{ (G_3, 0), (G_4, 0) \}, \{ (G_1, 1) \}, \{ (G_4, 1), (G_5, 0) \} \\ \{ \{ (G_1, 0), (G_2, 1), (G_3, 0), (G_4, 0), (G_5, 0) \}, \\ \{ (G_1, 1), (G_2, 1), (G_3, 0), (G_4, 0), (G_5, 0) \}, \\ \{ (G_1, 1), (G_2, 1), (G_3, 0), (G_4, 1), (G_5, 0) \} \} > \end{aligned}$$

Wyliczone wartości funkcji $attrval_{JEP}$ znajdują się w tabeli 8.3.2-1.

Atrybut	G ₁	G ₂	G ₃	G ₄	G ₅
$attrval_{JEP}$	0,500	1,000	0,875	0,500	0,625

Tab. 8.3.3-1. Wartości funkcji $attrval_{JEP}$ dla atrybutów tablicy TD_3

8.3.4. Porównanie miar istotności genów

Funkcje oceny istotności genów mogą być wykorzystane do dwóch podstawowych celów. Po pierwsze wskazują geny silnie związane z badaną hipotezą. Ich wartości mogą zostać wzajemnie porównane i odniesione do wyników uzyskanych metodami statystycznymi. Po drugie podobieństwo funkcji oceny zdefiniowanych na bazie reduktów i wzorców wyskakujących może wskazywać na związek pomiędzy tymi dwoma pojęciami.

W rozważaniach będziemy brali pod uwagę cztery funkcje oceny: $attrval_F$, $attrval_{\chi}$, $attrval_{RED}$, $attrval_{JEP}$. Ponieważ analiza dla wszystkich tych funkcji będzie przebiegała jednakowo, więc do przedstawienia sposobu porównywania będę stosował ogólne oznaczenia: $attrval_1$, $attrval_2$. Przy czym każda z tych funkcji reprezentuje jedną z funkcji badanych.

Do analizy podobieństwa ocen zastosujemy statystyczny pomiar korelacji ich rozkładów.

Wprowadźmy oznaczenia:

- $C = \{a_1, \dots, a_n\}$
- dla $i=1, \dots, n$: $x_i = attrval_1(a_i)$
- dla $i=1, \dots, n$: $y_i = attrval_2(a_i)$
- średnie wartości: $\bar{x} = \sum_{i=1}^n x_i$, $\bar{y} = \sum_{i=1}^n y_i$

Współczynnik korelacji wartości funkcji oceny dla atrybutów tablicy decyzyjnej jest

$$\text{wyrażony wzorem: } \rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Wzajemną zależność wartości obu funkcji można przedstawić na wykresie dwuwymiarowym, jako zbiór punktów $\{(x_i, y_i)\}_{i=1, \dots, n}$. Każdy z punktów wykresu odpowiada jednemu obiektowi.

Przykład

Tabela 8.3.4-1 przedstawia funkcje oceny atrybutów, które zostały wykorzystane do wartościowania genów w różnych etapach analizy tablicy decyzyjnej DT_1 . Atrybuty zostały uporządkowane zgodnie z malejącymi wartościami funkcji $attrval_F$. W tabeli 8.3.4-2 znajdują się podstawowe wielkości statystyczne opisujące rozkład wartości funkcji oceny, a w tabeli 8.3.4-3 – korelacje par rozkładów funkcji oceny. Na rysunku 8.3.4-2 został przedstawiony wykres korelacji rozkładu wartości funkcji $attrval_F$ i $attrval_{JEP}$.

Gen X	Próg podziału	$attrval_F(X)$	$attrval_{\chi}(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
G_2	5,330	1,000	1,000	1,000	1,000
G_3	0,780	0,905	0,577	0,833	0,875
G_5	3,345	0,893	0,775	0,500	0,625
G_1	4,755	0,821	0,577	0,333	0,500
G_4	5,065	0,662	0,500	0,333	0,500

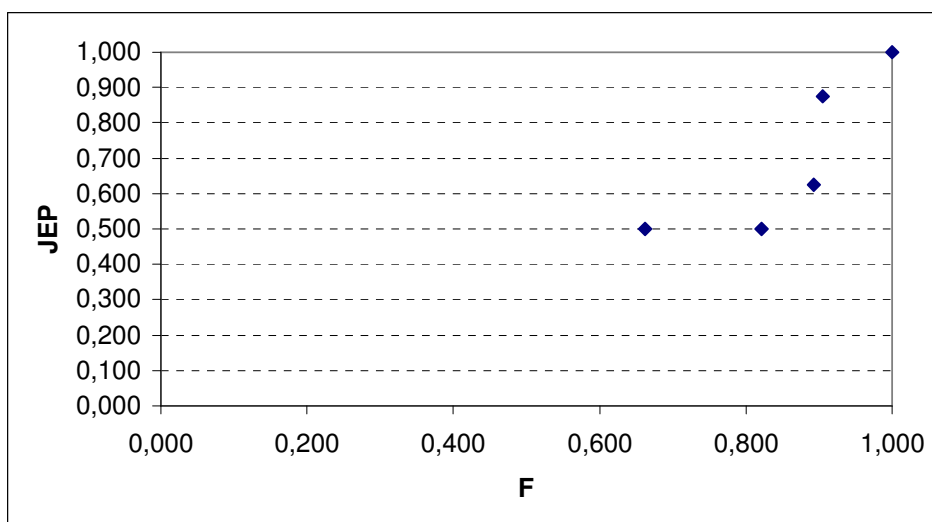
Tab.8.3.4-1. Dyskretyzacja i ocena genów znaczących

Wielkość	$attrval_F(X)$	$attrval_{\lambda}(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
Średnia	0,831	0,750	0,667	0,750
Wariancja	0,057	0,125	0,222	0,125
Odchylenie st.	0,239	0,354	0,471	0,354

Tab. 8.3.4-2. Statystyczne właściwości funkcji ocen genów

	$attrval_F(X)$	$attrval_{\lambda}(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
$attrval_F(X)$	1,000	0,814	0,829	0,829
$attrval_{\lambda}(X)$	0,814	1,000	0,692	0,692
$attrval_{RED}(X)$	0,829	0,692	1,000	1,000
$attrval_{JEP}(X)$	0,829	0,692	1,000	1,000

Tab. 8.3.4-3. Współczynniki korelacji wartości funkcji oceny



Rys.8.3.4-1. Wykres korelacji funkcji oceny: $attrval_F$ i $attrval_{JEP}$

8.4. Analiza postaci zbiorów wynikowych

Zbiór wszystkich reduktów i przestrzeń wzorców wyskakujących dla rzeczywistych zbiorów danych osiągają zwykle bardzo duże rozmiary. Ze względu na dużą wymiarowość zagadnienia i względnie małą liczbę obiektów nawet wstępna eliminacja części atrybutów nie rozwiązuje problemu dokładnej analizy otrzymanych wyników.

W pierwszej kolejności sprecyzujemy postać rezultatów. Najczęściej dysponujemy wyłącznie pewnym podzbiorem reduktów tablicy decyzyjnej o długości nie większej niż pewne ustalone ograniczenie. Zbiór ten otrzymujemy za pomocą algorytmu przybliżonego (rozdział 6.4.4). Zaznaczmy, że w tym wypadku dysponujemy bezpośrednio konkretnymi zbiorami atrybutów. Z kolei przestrzeń wzorców wyskakujących dla danego problemu jest wyznaczana w sposób dokładny, za pomocą algorytmu jep-producer (rozdział 7.4.6). Wynik jest zapisany w reprezentacji płotowej. W ten sposób dysponujemy bezpośrednio zbiorem najsilniejszych i najsłabszych wzorców wyskakujących z wybranej klasy rozpatrywanego zbioru danych.

Najlepszym sprawdzianem adekwatności prezentowanych metod do analizy ekspresji byłaby biologiczna weryfikacja sposobu funkcjonowania i znaczenia grup genów w badanym zjawisku. Ze względu na brak zaplecza biologicznego jedynym źródłem informacji są wyniki analiz różnych zespołów przeprowadzone na rozważanych w pracy zbiorach danych. Szczególnie cenne dla porównania są propozycje opisujące podejścia odmienne od tutaj przedstawionych.

Istotnym celem pracy jest porównanie pojęcia reduktu i wzorca wyskakującego. W tym celu pominięta zostanie dodatkowa informacja niesiona przez wzorzec na temat konkretnych wartości atrybutów i oba podejścia będą traktowane jako wskaźniki pewnych zbiorów atrybutów. Zbiory uzyskane w obu podejściach zostaną porównane w następujący sposób.

Zbadanie rozkładu podzbiorów różnej wielkości

Zaproponowana metoda jest oparta na tej samej idei, co podejście z rozdziału 8.3.4. Badana jest częstotliwość występowania podzbiorów określonej wielkości w każdej z obu analizowanych rodzin wynikowych. Ze względu na znaczne różnice w wielkości rodzin częstotliwość występowania podzbioru określonej wielkości musi być odpowiednio znormalizowana. W obu przypadkach wartość jest odnoszona do sumy wartości po wszystkich długościach zbiorów.

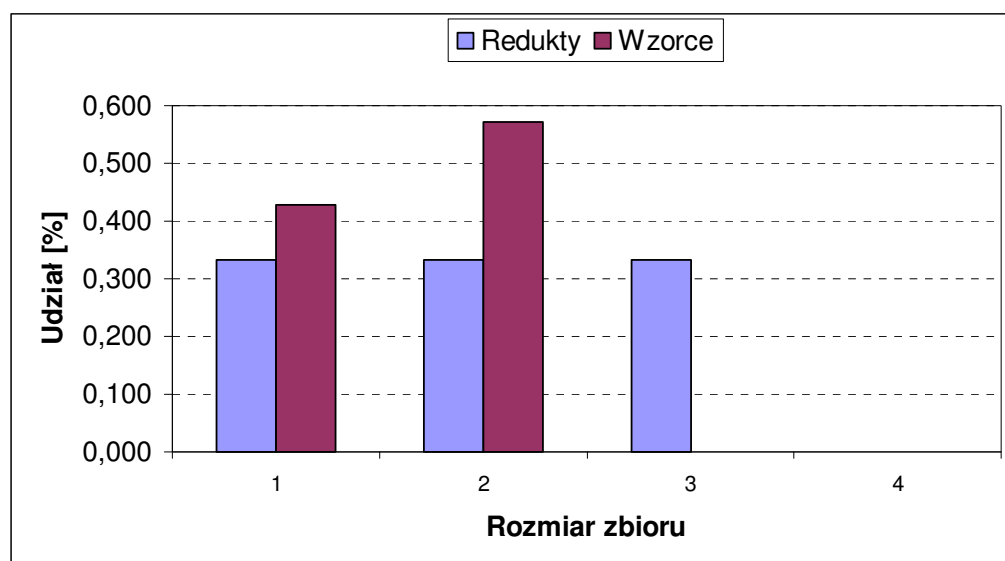
Wprowadźmy oznaczenia:

- R – wynikowy zbiór reduktów
- JF – lewy kraniec płotu przestrzeni wzorców wyskakujących
- dla $i=1,...,n$: $x_i = \frac{1}{|R|} |\{r \in R : |r| = i\}|$
- dla $i=1,...,n$: $y_i = \frac{1}{MAX} \sum_{J \in JF \wedge |J|=i} count_{D_2}(J)$, gdzie $MAX = \sum_{i=1,...,n} \left(\sum_{J \in JF \wedge |J|=i} count_{D_2}(J) \right)$

Biorąc pod uwagę powyższe oznaczenia współczynnik korelacji i wykres korelacji zdefiniowane są jak w rozdziale 8.3.4. Punkt (x_i, y_i) odpowiada podzbirowi atrybutów o długości i .

Przykład

Rysunek 8.4-1 zawiera histogramy odpowiadające z seriom danych dla ciągów: $\{x_i\}$ i $\{y_i\}$. Tak mała liczba reduktów i wzorców wyskakujących nie pozwala na wysnucie interesujących wniosków.



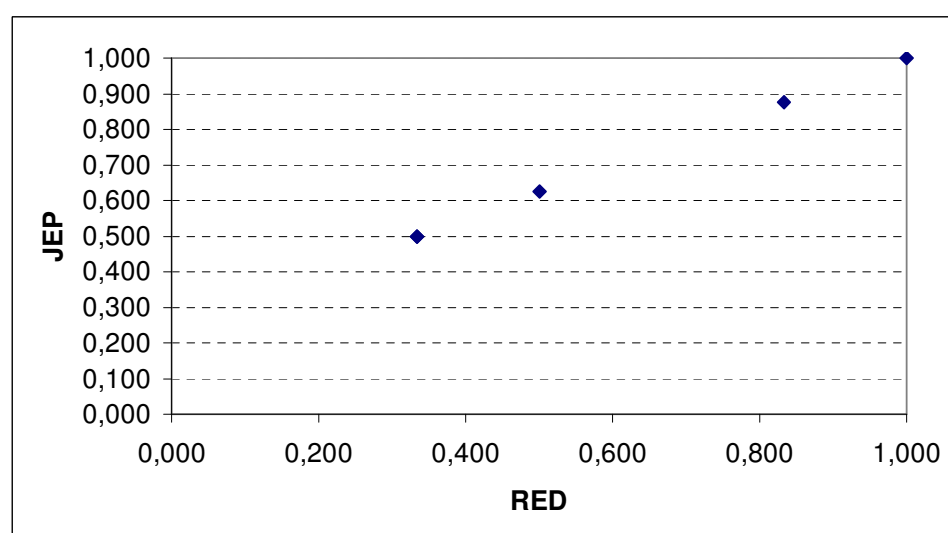
Rys. 8.4-1. Histogramy dla podziału reduktów i wzorców na klasy według rozmiarów

Zbadanie częstotliwości występowania atrybutów

Badanie istotności pojedynczego atrybutu za pomocą metod z rozdziału 8.3.4 może służyć również porównaniu postaci zbiorów reduktów i przestrzeni wzorców wyskakujących. Ukonkretniając, porównane zostaną miary: $attrval_{RED}$ i $attrval_{JEP}$. Analiza obejmie obliczenie współczynnika korelacji i sporządzenie wykresu korelacji.

Przykład

Wykres korelacji na rysunku 8.4-2 potwierdza pełną korelację rozkładu wartości funkcji $attrval_{RED}$ i $attrval_{JEP}$ (tabela 8.3.4-3). Należy podkreślić, że zależność ta jest cechą dla badanej tablicy decyzyjnej DT_3 i nie musi być regułą w ogólności.



Rys. 8.4-2. Wykres korelacji funkcji oceny: $attrval_{RED}$ i $attrval_{JEP}$

9. Wyniki eksperymentalne

9.1. Sposób prowadzenia analizy

W pracy przedstawiono ścieżkę analizy danych otrzymanych w wyniku eksperymentu z wykorzystaniem mikromacierzy. Zakłada ona, że zbiór wejściowy przeszedł przetwarzanie wstępne i jest zapisany w postaci tablicy decyzyjnej. Badania składają się z oszacowania wartości genów za pomocą metody analizy wariancji, wyboru pewnego zbioru genów znaczących, dyskretyzacji nowej tablicy decyzyjnej i ostatecznie znalezieniu zbioru reduktów oraz przestrzeni wzorców wyskakujących. Dla niektórych etapów zaproponowane zostały różne alternatywne opcje. Wszystkie etapy ścieżki przetwarzania można realizować za pomocą specjalnie przygotowanej aplikacji Microarray Analyser.

W analizie rzeczywistych danych zbiory wymagały wstępnego przetworzenia (np. uzupełnienia wartości brakujących) oraz konwersji do formatu akceptowanego przez program. Modyfikacje zostały skomentowane oddzielnie dla każdego zbioru.

Do szacowania wartości genów wykorzystana została, często pojawiająca się w literaturze, metoda analizy wariancji. Zbiór znaczący definiowany był jako zbiór rankingowo znaczący z odpowiednio dobranym parametrem λ .

Duża część informacji jest tracona na etapie dyskretyzacji. Z tego względu wykorzystamy metodę zstępującą opartą na analizie entropii, gdyż uwzględnia ona hipotezę docelową przy wyznaczaniu podziałów zbiorów wartości atrybutów. Metoda może prowadzić do dodatkowej redukcji zbioru atrybutów. Z dalszych rozważań eliminowane są atrybuty zdyskretyzowane do jednego przedziału.

Do wyszukiwania reduktów zostały wykorzystane: algorytm dokładny i przybliżony evolutionary-reduct-finder opisany w rozdziale 6.4.4. Algorytm i parametry wejściowe, w szczególności zakres przeglądanej przestrzeni, zostały dobrane do konkretnego zbioru danych. Decyzja wynikała ze względów efektywnościowych. Ze względu na rozmiary zbioru reduktów w pracy zamieszczono zbiory wszystkich najkrótszych i najdłuższych reduktów.

Przestrzeń wzorców wyskakujących została wyznaczona za pomocą algorytmu jep-producer przedstawionego w rozdziale 7.4.6. Ze względu na duży rozmiar płotu przestrzeni nie został on podany w całości.

Analiza uzyskanych wyników przebiegała zgodnie z metodami przedstawionymi w rozdziale 8. Obejmowała ona porównanie miar znaczenia atrybutów opartych na analizie wariancji, statystyce χ^2 , reduktach i wzorcach wyskakujących.

9.2. Wyniki eksperymentów

W pracy zostały przeanalizowane dwa zbiory danych pochodzące z opublikowanych eksperymentów. Spośród znalezionych źródeł wybór padł na eksperymenty najszczegółowiej udokumentowane, dla których dostępne były wstępnie przetworzone dane wynikowe.

9.2.1. Zbiór danych „Ontario”

Charakterystyka zbioru

Eksperyment dotyczy klasyfikacji pacjentów na chorych na rakach i zdrowych na podstawie badania ekspresji genów. Dokładny opis zawarty jest w pracy [46], a wyniki są dostępne na stronie [55]. W celach ilustracyjnych przebieg omawianego doświadczenia i postać danych wynikowych zostały przedstawione w rozdziale 2.9.

Do analizy został wybrany zbiór danych przygotowany przez autorów, który ma następujące cechy:

- Badane poziomy zostały odniesione do poziomów kontrolnych
- Została przeprowadzona normalizacja dla całego eksperymentu
- Geny o znacząco niekompletnych informacjach o poziomach ekspresji na przestrzeni badanych przypadków zostały wyeliminowane z dalszej analizy
- W macierzy ekspresji występują brakujące wartości
- Plik w formacie XLS

Dane na temat klas decyzyjnych przyporządkowanych przypadkom zostały pobrane z oddzielnego pliku opisującego pacjentów.

Przeprowadzenie analizy za pomocą programu wymagało następującego przetwarzania wstępnego:

- Uzupełnienie wartości brakujących ustaloną wartością równą średniej arytmetycznej z dostępnych poziomów ekspresji danego genu dla poszczególnych przypadków
- Dostosowania plików do postaci akceptowanej przez program

Wielkość zbioru danych „Ontario”		
Liczba przykładów	Liczba genów	Liczba klas
39	2880	2

Tab. 9.2.1-1. Wielkość zbioru danych „Ontario”

Zastosowane metody

Dla wszystkich genów została wyliczona wartość funkcji attrval_F , a następnie został wybrany zbiór 0,4;1-rankingowo znaczący zawierający 249 genów. Celowo obrano niski próg, aby badania można było przeprowadzić na większej liczbie genów.

Zbiór został zdyskretyzowany zstępująco. Duża część genów została zdyskretyzowana do jednego przedziału. Aby geny te nie zaciemniały obrazu dalszej analizy, zostały one usunięte z tablicy decyzyjnej. Zredukowana tablica zawierała 45 genów.

Ze względów efektywnościowych redukt zostały wyznaczone za pomocą metody przybliżonej evolutionary-reduct-finder z parametrami zgromadzonymi w tabeli 9.2.1-2.

Parametr	Wartość
seed	123
crossoverRate	0.3
mutationRate	0.2
populationSize	100
iterationCount	100
sucesionRate	0.2
bestSetSize	500
minAttributeCount	3
maxAttributeCount	20

Tab. 9.2.1-2. Parametry algorytmu evolutionary-reduct-finder

Następnie w sposób dokładny została wyznaczona przestrzeń wzorców wyskakujących.

Ocena genów

Nr	Gen X	Próg podziału	$attrval_F(X)$	$attrval_{\chi}(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
1	121524	0,668	1,000	0,626	0,015	0,219
2	116692	0,000	0,929	0,622	0,011	0,229
3	193109	0,453	0,859	0,568	0,007	0,432
4	200058	0,000	0,844	0,580	0,006	0,289
5	47345	0,418	0,823	0,592	0,006	0,177
6	198612	0,573	0,816	0,615	0,018	0,352
7	50711	0,000	0,807	0,568	0,001	0,330
8	42837	0,473	0,789	0,592	0,022	0,212
9	21593	0,523	0,775	0,580	0,099	0,273
10	113699	0,538	0,755	0,743	0,065	0,186
11	191827	0,000	0,755	0,520	0,046	0,414
12	121479	0,473	0,739	0,568	0,068	0,234
13	31795	0,577	0,733	0,743	0,126	0,219
14	41816	0,523	0,733	0,687	0,136	0,161
15	110608	0,841	0,724	0,539	0,034	0,218
16	212856	0,000	0,712	0,622	0,183	0,254
17	446652	0,000	0,707	0,447	0,048	0,106
18	49602	0,488	0,689	0,643	0,064	0,256
19	30295	0,490	0,673	0,626	0,209	0,341
20	321970	0,610	0,654	0,539	1,000	1,000
21	204656	0,530	0,644	0,592	0,018	0,210
22	33408	0,000	0,644	0,539	0,081	0,214
23	119800	0,553	0,612	0,527	0,042	0,005
24	294136	0,000	0,606	0,496	0,010	0,004
25	510651	0,785	0,600	0,413	0,111	0,099
26	205602	0,000	0,600	0,539	0,041	0,147
27	37087	0,000	0,599	0,568	0,104	0,258
28	167296	0,000	0,572	0,615	0,203	0,293
29	207793	0,480	0,565	0,580	0,189	0,210

30	115012	0,000	0,534	0,539	0,048	0,199
31	430207	0,000	0,520	0,496	0,012	0,004
32	22984	0,000	0,518	0,568	0,184	0,222
33	31287	0,000	0,515	0,622	0,211	0,279
34	39204	0,000	0,505	0,617	0,162	0,264
35	429621	0,000	0,503	0,527	0,064	0,005
36	122945	0,643	0,495	0,568	0,117	0,330
37	114082	0,000	0,466	0,568	0,235	0,526
38	30111	0,000	0,448	0,626	0,203	0,288
39	140870	0,000	0,442	0,539	0,080	0,182
40	197637	2,060	0,432	0,693	0,176	0,339
41	202704	0,000	0,430	0,626	0,229	0,220
42	112608	0,000	0,427	0,622	0,264	0,136
43	211255	0,503	0,418	0,539	0,024	0,055
44	376316	0,490	0,406	0,496	0,071	0,004
45	174237	0,000	0,406	0,617	0,220	0,267

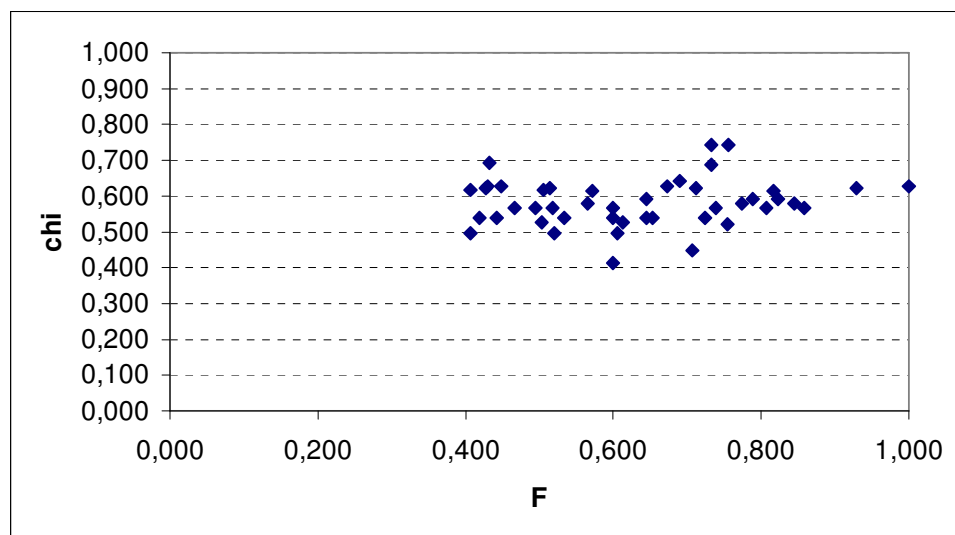
Tab. 9.2.1-3. Dyskretyzacja i ocena genów znaczących

Wielkość	$attrval_F(X)$	$attrval_\chi(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
Średnia	0,703	0,622	0,117	0,243
Wariancja	0,176	0,000	0,021	0,001
Odchylenie st.	0,420	0,007	0,146	0,034

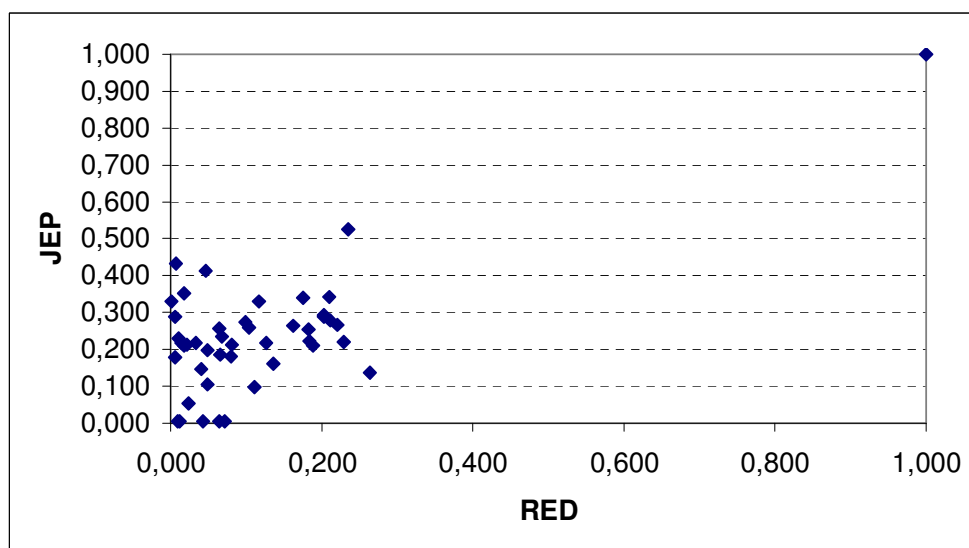
Tab. 9.2.1-4. Statystyczne właściwości funkcji ocen genów

	$attrval_F(X)$	$attrval_\chi(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
$attrval_F(X)$	1,000	0,157	-0,275	0,170
$attrval_\chi(X)$	0,157	1,000	0,091	0,199
$attrval_{RED}(X)$	-0,275	0,091	1,000	0,716
$attrval_{JEP}(X)$	0,170	0,199	0,716	1,000

Tab. 9.2.1-5. Współczynniki korelacji wartości funkcji oceny



Rys. 9.2.1-1. Wykres korelacji funkcji oceny: $attrval_F$ i $attrval_\chi$



Rys. 9.2.1-2. Wykres korelacji funkcji oceny: $attrval_{RED}$ i $attrval_{JEP}$

Zbiór reduktów i przestrzeń wzorców wyskakujących

Rozmiar reduktu	Liczba reduktów
3	6
4	243
5	480
6	125
7	23
8	5
9	1
Suma	883

Tab. 9.2.1-6. Rozmiary wyznaczonych reduktów

Rozmiar	Redukt	Rozmiar	Redukt
3	{5, 13, 20}	8	{4, 14, 19, 25, 28, 43, 44, 45}
	{8, 10, 20}		{4, 14, 19, 28, 35, 38, 43, 45}
	{10, 13, 20}		{11, 14, 19, 27, 29, 33, 35, 38}
	{10, 20, 22}		{14, 16, 17, 19, 28, 34, 43, 45}
	{13, 18, 20}		{14, 16, 19, 25, 28, 34, 43, 45}
	{13, 20, 40}		
		9	{9, 11, 14, 19, 24, 31, 33, 36, 37}

Tab. 9.2.1-7. Zbiór reduktów o rozmiarach 3, 8 i 9

$count_{D_2}(X)$	$supp_{D_2}(X)$	Liczba wzorców
24	1,000	12
23	0,958	114
22	0,917	93
21	0,875	46
20	0,833	41
19	0,792	52
18	0,750	23
17	0,708	11
16	0,667	4
15	0,625	8
14	0,583	7
13	0,542	9
12	0,500	30
11	0,458	20
10	0,417	17
9	0,375	7
8	0,333	3
7	0,292	12
6	0,250	14
5	0,208	91
4	0,167	1195

3	0,125	1108
2	0,083	1649
1	0,042	4868
	Suma	9434

Tab. 9.2.1-8. Wzorce w lewym krańcu płotu przestrzeni wzorców wyskakujących

$count_{D_2}(X)$	$supp_{D_2}(X)$	Liczba wzorców
3	0,125	1
2	0,083	1
1	0,042	17
	Suma	19

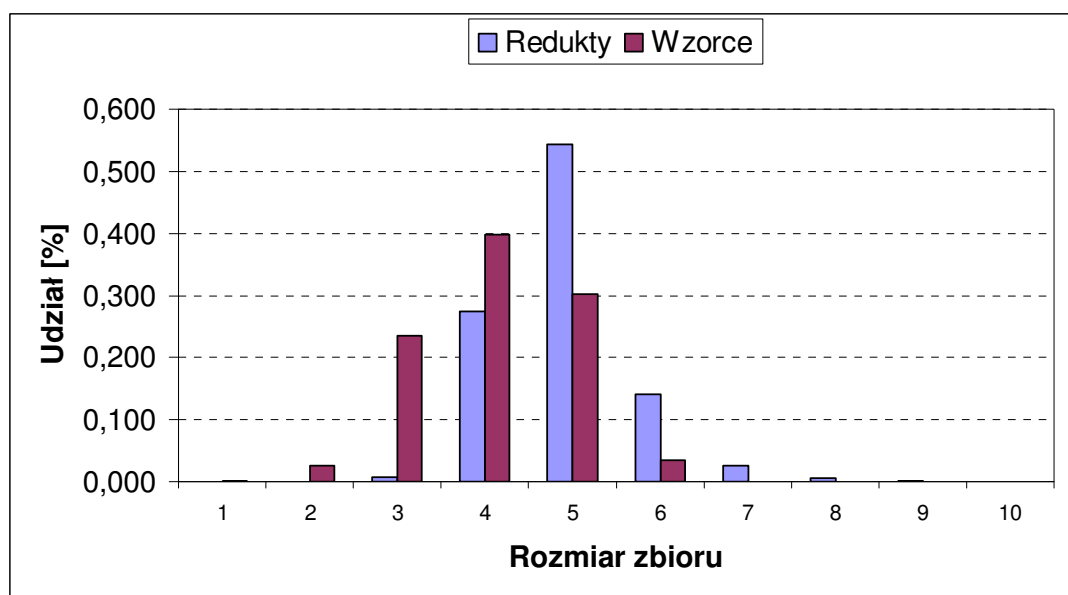
Tab. 9.2.1-9. Wzorce w prawym krańcu płotu przestrzeni wzorców wyskakujących

Wzorec X	$count_{D_2}(X)$	$supp_{D_2}(X)$
$\{(8,1),(10,1),(20,0)\}$	24	1,000
$\{(8,1),(15,1),(20,0),(40,1)\}$	24	1,000
$\{(8,1),(20,0),(30,1),(40,1)\}$	24	1,000
$\{(8,1),(20,0),(39,1),(40,1)\}$	24	1,000
$\{(10,1),(13,1),(20,0)\}$	24	1,000
$\{(10,1),(20,0),(22,1)\}$	24	1,000
$\{(13,1),(18,1),(20,0)\}$	24	1,000
$\{(13,1),(20,0),(40,1)\}$	24	1,000
$\{(15,1),(20,0),(22,1),(40,1)\}$	24	1,000
$\{(20,0),(22,1),(30,1),(40,1)\}$	24	1,000
$\{(20,0),(22,1),(39,1),(40,1)\}$	24	1,000
$\{(5,1),(13,1),(20,0)\}$	24	1,000

Tab. 9.2.1-10. Wzorce wyskakujące lewego krańca płotu o najwyższej ekspresji

Rozmiar	Udział reduktów	Udział wzorców
1	0,000	0,002
2	0,000	0,026
3	0,007	0,236
4	0,275	0,398
5	0,544	0,303
6	0,142	0,035
7	0,026	0,000
8	0,006	0,000
9	0,001	0,000
10	0,000	0,000

Tab. 9.2.1-11. Podział reduktów i wzorców na klasy według rozmiarów



Rys. 9.2.1-3. Histogramy dla podziału reduktów i wzorców na klasy według rozmiarów

Ustosunkowanie do innych prac

Rozważyliśmy wyniki pracy [46] opublikowanej przez autorów eksperymentu. Autorzy stosowali statystyczny model Cox'a do znalezienia genów znaczących oraz grupowanie hierarchiczne przykładów do wyznaczenia grup genów o różnych profilach ekspresji. Wyniki oceny statystycznej znajdują się w tabeli 9.2.1-12. Podane przez autorów numery genów zostały przypisane do identyfikatorów na arkuszu. Trzy geny nie zostały odnalezione w zbiorach wynikowych.

Identyfikator w arkuszu	Numer GenBank	Identyfikator Unigene	Opis genu	Wartość p
Brak	AB001105	Hs.272891	Hippocalcin-like protein 4 (HPCAL4)	0,0006
52648	H29772	Hs.32501	predicted protein	0,0006
44075	H06235	Hs.12294	predicted protein	0,0007
32555	H14691	Hs.159161	Rho GDP dissociation inhibitor (GDI) α	0,0007
50416	H17978	Hs.21814	Interleukin 20 receptor, α	0,0011
121145	T97149	Hs.83575	KIAA1533 predicted protein	0,0012
21649	T65437	Hs.268053	KIAA0029 protein	0,0012
206375	H58462	Hs.14317	NOLA3 Nucleolar protein family A, member 3	0,0014
Brak	R40106	Hs.286145	RNA polymerase II holoenzyme component SRB7	0,0016
146858	R80974	Hs.75188	Wee1 + (<i>S. pombe</i>) homolog protein kinase	0,0017

109111	AF020760	Hs.115721	serine protease 25 (PRSS25)	0,0017
147235	R80770	Hs.30819	C40 Hypothetical protein	0,0018
156097	R73128	Hs.73931	HLA-DQB1 MHC	0,0019
Brak	AF201940	Hs.283740	DC6 predicted protein	0,0020
34850	R20380	Hs.12758	predicted protein	0,0021

Tab. 9.2.1-12. Geny znaczące ($p \leq 0,0023$) zgodnie ze statystycznym modelem Cox'a

9.2.2. Zbiór danych „Princeton”

Charakterystyka zbioru

Eksperyment został przeprowadzony przez grupę badawczą na Princeton University. Dokładny opis zamieszczony jest w pracy [1], a dane dostępne są na stronie [57].

Badaniu zostało poddane 62 tkanki pochodzących z ludzkiej okrężnicy. Tkanki były różnych typów: 44 zrakowaciałe i 22 zdrowe. Do pomiaru ekspresji wykorzystano macierze oligonukleotydowe zawierające około 6600 rodzajów sekwencji genetycznych. Każda sekwencja była reprezentowana przez około 20 par oligonukleotydów. Otrzymane dane zostały oczyszczone i znormalizowane. Spośród badanych sekwencji wybrano 2000 genów o największych minimalnych intensywnościach ekspresji pomiędzy wszystkimi przykładami. Pozostałe sekwencje zostały uznane za nieistotne.

Autorzy skupili się na analizie danych bez nadzoru. Wykorzystali algorytm grupowania do wykrycia genów o wysoce skorelowanych wzorcach ekspresji oraz do badania tkanek o podobnym stanie. Informacje na temat tkanek zostały odniesione do znanego, faktycznego stanu tkanki.

Analiza została przeprowadzona na jedynym udostępnionym publicznie przez autorów zbiorze danych.

- Informacje na temat każdego genu zostały uwzględnione
- Została przeprowadzona normalizacja dla całego eksperymentu
- Plik w formacie tekstowym

Zbiór danych został dostosowany do postaci wymaganej przez program. W szczególności zmodyfikowany został sposób zapisu liczb rzeczywistych.

Wielkość zbioru danych „Princeton”		
Liczba przykładów	Liczba genów	Liczba klas
62	2000	2

Tab. 9.2.2-1. Wielkość zbioru danych „Princeton”

Zbiór był analizowany w pracach [29,30,31] pod kątem wyznaczania przestrzeni wzorców wyskakujących i oceny ich znaczenia dla dziedziny.

Zastosowane metody

Dla wszystkich genów została wyliczona wartość funkcji $attrval_F$, a następnie został wybrany zbiór 0,552;1-rankingowo znaczący. Próg został tak wybrany, aby zbiór znaczący zawierał 35 geny. Liczba ta pozwala na dokładne wyznaczenie zbioru reduktów. Zbiór został zdyskretyzowany zstępująco, przy czym zbiory wartości wszystkich genów zostały podzielone na 2 przedziały. Następnie w sposób dokładny zostały wyznaczone: zbiór reduktów i przestrzeń wzorców wyskakujących.

Ocena genów

Nr	Gen X	Próg podziału	$attrval_F(X)$	$attrval_{\chi}(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
1	M63391	1696,230	1,000	0,714	0,266	0,365
2	M76378	842,305	0,924	0,661	0,392	0,020
3	R87126	379,390	0,910	0,696	0,388	0,918
4	J02854	735,805	0,906	0,648	0,452	0,176
5	M76378	1048,375	0,895	0,641	0,263	0,409
6	M76378	1136,750	0,877	0,641	0,263	0,409
7	Z50753	196,125	0,811	0,474	1,000	0,935
8	T92451	845,705	0,802	0,577	0,470	0,679
9	U25138	186,195	0,723	0,567	0,277	0,385
10	H08393	84,880	0,700	0,628	0,498	0,622
11	T71025	1695,985	0,690	0,577	0,407	0,582
12	H43887	334,015	0,663	0,583	0,473	0,578
13	J05032	88,905	0,655	0,642	0,435	0,450
14	X63629	84,195	0,644	0,588	0,885	0,838
15	H40095	400,035	0,636	0,614	0,414	0,538
16	M26697	1905,800	0,627	0,437	0,602	0,410
17	X86693	319,475	0,624	0,527	0,593	0,650
18	X12671	229,995	0,610	0,626	0,467	0,018
19	M36634	58,505	0,606	0,474	0,593	0,704
20	M22382	447,045	0,602	0,642	0,392	0,460
21	T95018	6225,980	0,589	0,401	0,229	0,272
22	T57619	2597,850	0,587	0,493	0,509	0,448
23	T56604	456,370	0,584	0,456	0,176	0,259
24	T60155	1798,655	0,580	0,564	0,432	0,632
25	U09564	0,000	0,573	0,577	0,469	0,482
26	T47377	120,385	0,572	0,538	0,746	1,000
27	R84411	629,440	0,569	0,493	0,864	0,764
28	T60778	0,000	0,566	0,529	0,389	0,542
29	X14958	735,570	0,563	0,516	0,557	0,471
30	T62947	0,000	0,560	0,437	0,301	0,432
31	H55916	816,330	0,553	0,382	0,418	0,539

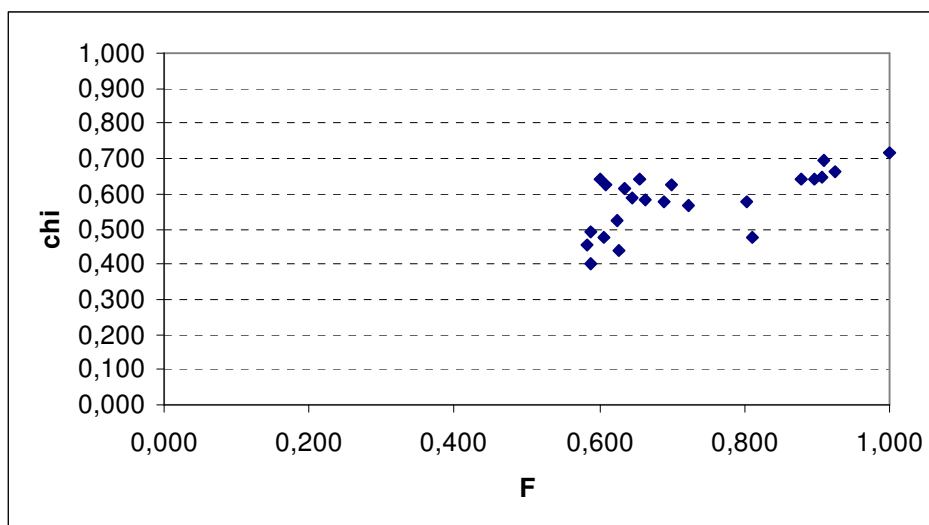
Tab. 9.2.2-2. Dyskretyzacja i ocena genów znaczących

Wielkość	$attrval_F(X)$	$attrval_\chi(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
Średnia	0,777	0,548	0,342	0,452
Wariancja	0,100	0,055	0,011	0,015
Odchylenie st.	0,316	0,235	0,107	0,123

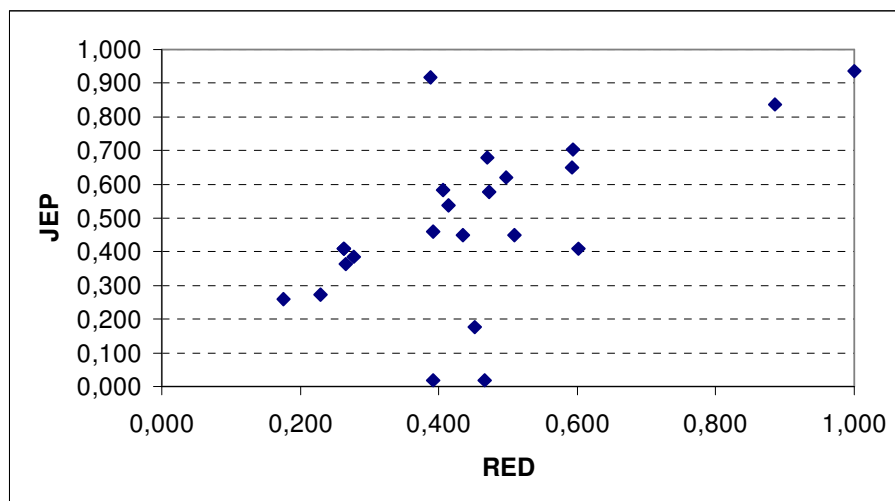
Tab. 9.2.2-3. Statystyczne właściwości funkcji ocen genów

	$attrval_F(X)$	$attrval_\chi(X)$	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
$attrval_F(X)$	1,000	0,666	-0,202	-0,147
$attrval_\chi(X)$	0,666	1,000	-0,200	-0,130
$attrval_{RED}(X)$	-0,202	-0,200	1,000	0,631
$attrval_{JEP}(X)$	-0,147	-0,130	0,631	1,000

Tab. 9.2.2-4. Współczynniki korelacji wartości funkcji oceny



Rys. 9.2.2-1. Wykres korelacji funkcji oceny: $attrval_F$ i $attrval_\chi$



Rys. 9.2.2-2. Wykres korelacji funkcji oceny: $attrval_{RED}$ i $attrval_{JEP}$

Zbiór reduktów i przestrzeń wzorców wyskakujących

Rozmiar reduktu	Liczba reduktów
5	2
6	213
7	1466
8	3884
9	8356
10	7334
11	1074
12	186
Suma	22515

Tab. 9.2.1-5. Rozmiary wyznaczonych reduktów

Rozmiar	Redukt
6	{1, 2, 4, 10, 19, 26}
	{1, 2, 10, 15, 19, 26}
	{1, 2, 15, 19, 22, 26}
	{1, 3, 4, 10, 19, 26}
	{1, 3, 7, 19, 26, 27}
	{1, 3, 10, 13, 19, 26}
	{1, 3, 10, 14, 19, 26}
	{1, 3, 10, 15, 19, 26}
	{1, 3, 10, 19, 26, 27}
	{1, 3, 10, 19, 26, 29}
	{1, 3, 10, 19, 26, 30}
	{1, 3, 10, 19, 26, 31}
	{1, 3, 19, 22, 26, 27}
	{1, 3, 19, 22, 26, 30}
	{1, 3, 19, 22, 26, 31}
	{1, 4, 10, 13, 19, 26}
	{1, 4, 10, 15, 18, 19}
6	{1, 4, 10, 15, 19, 26}
	{1, 4, 10, 16, 18, 29}
	{1, 4, 10, 18, 19, 20}
	{1, 4, 10, 18, 19, 25}
	{1, 4, 10, 18, 19, 26}
	{1, 4, 10, 18, 19, 29}
	{1, 4, 10, 19, 20, 26}
	{1, 4, 10, 19, 25, 26}
	{1, 10, 13, 19, 24, 26}
	{1, 10, 13, 19, 26, 28}
	{1, 10, 14, 15, 17, 19}
	{1, 10, 14, 15, 19, 26}
	{1, 10, 15, 18, 19, 24}

Tab. 9.2.1-6. Zbiór wybranych reduktów o długości 6

$count_{D2}(X)$	$supp_{D2}(X)$	Liczba wzorców
17	0,810	4
16	0,762	19
15	0,714	8
14	0,667	10
13	0,619	32
12	0,571	37
11	0,524	32
10	0,476	43
9	0,429	47
8	0,381	44
7	0,333	68
6	0,286	86
5	0,238	64
4	0,190	206
3	0,143	251
2	0,095	830
1	0,048	1993
Suma		768

Tab. 9.2.1-7. Wzorce w lewym krańcu płotu przestrzeni wzorców wyskakujących

$count_{D2}(X)$	$supp_{D2}(X)$	Liczba wzorców
2	0,095	2
1	0,048	18
Suma		20

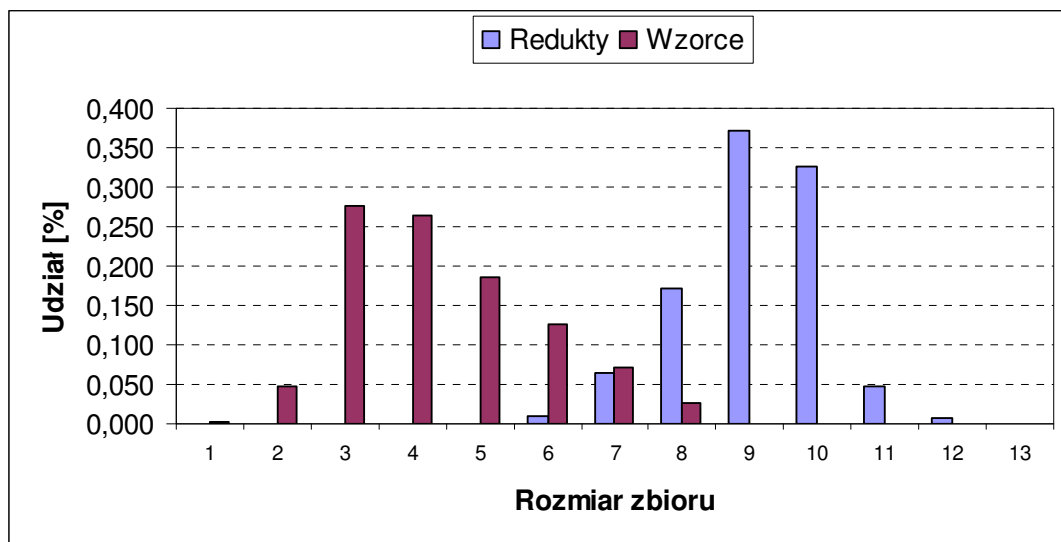
Tab. 9.2.1-8. Wzorce w prawym krańcu płotu przestrzeni wzorców wyskakujących

Wzorzec X	$count_{D2}(X)$	$supp_{D2}(X)$
$\{(15,0),(19,1),(26,0)\}$	17	0,810
$\{(3,1),(26,0),(27,0)\}$	17	0,810
$\{(3,1),(26,0),(30,0)\}$	17	0,810
$\{(3,1),(26,0),(31,0)\}$	17	0,810

Tab. 9.2.1-9. Wzorce wyskakujące lewego krańca płotu o najwyższej ekspresji

Rozmiar	Udział reduktów	Udział wzorców
1	0,000	0,002
2	0,000	0,049
3	0,000	0,275
4	0,000	0,265
5	0,000	0,185
6	0,009	0,126
7	0,065	0,072
8	0,173	0,027
9	0,371	0,000
10	0,326	0,000
11	0,048	0,000
12	0,008	0,000
13	0,000	0,000

Tab. 9.2.2-10. Podział reduktów i wzorców na klasy według rozmiarów



Rys. 9.2.2-3. Histogramy dla podziału reduktów i wzorców na klasy według rozmiarów

Ustosunkowanie do innych prac

W pracach [29,30,31] opisana została analiza rozważanego zbioru danych. Autorzy zastosowali dyskretyzację zstępującą do całego zakresu badanych genów. Następnie wybrali 135 genów, dla których metoda wyznaczyła co najmniej dwuelementowy podział. Spośród otrzymanych genów zostało następnie wybranych 35 genów, których podział otrzymał charakteryzował się największą entropią. Dla tak zredukowanej tablicy decyzyjnej zostały wyznaczone wzorce wyskakujące. W pracach [30,31] etap dyskretyzacji został przeprowadzony nieprawidłowo, co doprowadziło do błędnych wyników końcowych. Poprawne rezultaty zawiera wyłącznie korekta [29].

Nr według autorów	Gen X	Próg podziału	$attrval_F(X)$	Rozważany
1	M26383	59,83	0,353	Nie
2	M63391	1696,22	1,000	Tak
3	R87126	379,38	0,910	Tak
4	M76378	842,30	0,924	Tak
5	H08393	84,87	0,700	Tak
6	X12671	229,99	0,610	Tak
7	R36977	274,96	0,501	Nie
8	J02854	735,80	0,906	Tak
9	M22382	447,04	0,602	Tak
10	J05032	88,90	0,655	Tak
11	M76378	1048,37	0,895	Tak
12	M76378	1136,74	0,877	Tak
13	M16937	390,44	0,311	Nie
14	H40095	400,03	0,636	Tak
15	U30825	288,99	0,525	Nie
16	H43887	334,01	0,663	Tak
17	H51015	84,19	0,001	Nie
18	X57206	417,30	0,141	Nie
19	R10066	494,17	0,485	Nie
20	T96873	75,42	0,340	Nie
21	T57619	2597,85	0,587	Tak
22	R84411	735,57	0,569	Tak
23	U21090	232,74	0,432	Nie
24	U32519	87,58	0,501	Nie
25	T71025	1695,98	0,690	Tak
26	T92451	845,70	0,802	Tak
27	U09564	120,38	0,573	Tak
28	H40560	913,77	0,533	Nie
29	T47377	629,44	0,572	Tak
30	X53586	121,91	0,361	Nie
31	U25138	186,19	0,723	Tak
32	T60155	1798,65	0,580	Tak
33	H55758	1453,15	0,537	Nie
34	Z50753	196,12	0,811	Tak
35	U09587	486,17	0,528	Nie

Tab. 9.2.2-11. Geny analizowane w pracy [25] wybrane według kryterium entropii, uzupełnione o wartość funkcji oceny $attrval_F$ wyliczoną w niniejszej analizie

Ocena według entropii jest różna od oceny zgodnie ze statystyką F . Oba podejścia wskazały tych samych 21 genów. Kryterium entropii podało 14 genów, które zostały nisko ocenione zgodnie z przyjętą tu funkcją oceny $attrval_F$ i przez to wyeliminowane z dalszej analizy.

9.3. Omówienie wyników

9.3.1. Ocena pojedynczych genów

Rozdział dotyczy wyników otrzymanych przy zastosowaniu różnych funkcji oceny genów. Wyniki odniesione są do ocen dostępnych w publikacjach.

Dla zbioru „Ontario” analiza wariancji wskazała 9 sekwencji o $attrval_F(X) > 0,77$. Opisy pochodzą z arkuszy wyników sporządzonych przez autorów.

Gen X	Opis	$attrval_F(X)$
121524	121524:Data not found:not available	1,000
116692	116692:Data not found:not available	0,929
193109	193109:Data not found:not available	0,859
200058	200058:Data not found:not available	0,844
47345	47345:Data not found:not available	0,823
198612	198612:Hs.35406:not available	0,816
50711	50711:Hs.31500:not available	0,807
42837	42837:Hs.183475:not available	0,789
21593	21593:Hs.290263:not available	0,775

Tab. 9.3.1-1. Geny wskazane przez funkcję oceny $attrval_F$

Pięć wskazanych sekwencji nie zostało poprawnie zidentyfikowanych przez autorów. Cztery pozostałe wskazują na konkretne geny opisane w odpowiednich bazach sekwencji.

Funkcje $attrval_{RED}$ i $attrval_{JEP}$ najwyżej oceniły tę samą sekwencję odpowiadającą konkretnemu genowi. Może to przemawiać za jej istotnością dla powstawania nowotworów.

Gen X	Opis	$attrval_{RED}(X)$	$attrval_{JEP}(X)$
321970	321970:Hs.268231:not available	1,000	1,000

Tab. 9.3.1-2. Gen wskazany przez funkcje oceny $attrval_{RED}$ i $attrval_{JEP}$

Żaden z genów wskazanych przez model statystyczny Cox’a, wykorzystany przez autorów eksperymentu (tabela 9.2.1-12), nie pokrywa się z genami zbioru znaczącego rozważanego w naszej analizie.

Dla zbioru „Princeton” analiza wariancji wskazała 9 sekwencji $attrval_F(X) > 0,8$. Opisy pochodzą z arkuszy wyników sporządzonych przez autorów.

Gen X	Opis	$attrval_F(X)$
M63391	Human desmin gene, complete cds.	1,000
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0,924
R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)	0,910
J02854	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE	0,906

	ISOFORM (HUMAN);contains element TAR1 repetitive element	
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0,895
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0,877
Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor	0,811
T92451	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN)	0,802

Tab. 9.3.1-1. Geny wskazane przez funkcję oceny attrval_F

Zbiory znaczące ustalone za pomocą stosowanej w pracy analizy wariancji oraz wykorzystanej przez autorów prac [29,30,31] analizy entropii pokryły się w przypadku 21 genów. W szczególności obie oceny wskazały 3 replikanty tej samej sekwencji M76378.

Rozbieżność wyników wskazuje na fakt, jak bardzo istotny jest etap oceny i wyboru genów. Przykłady pokazały, że różne modele statystyczne, i analiza entropii podziału mogą prowadzić do zupełnie innych wyników. Dla poprawnych badań ważne jest, aby znaleźć uzasadnienie stosowania jednego podejścia lub rodziny podejść. Jeśli etap ten nie został przeprowadzony odpowiednio dalsza analiza może prowadzić do błędnych wniosków.

9.3.2. Grupy genów współzależnych

Grupy genów współzależnych dla obu zbiorów danych zostały wyznaczone za pomocą reduktów i wzorców wyskakujących.

W przypadku zbioru „Ontario” mamy następujące redukty o długości 6 (tabela 9.3.2-1). Zbiór został wyznaczony metodą przybliżoną.

Redukt zgodnie z numeracją genów	Redukt zgodnie z nazwami genów
{5, 13, 20}	{47345, 31795, 321970}
{8, 10, 20}	{42837, 113699, 321970}
{10, 13, 20}	{113699, 31795, 321970}
{10, 20, 22}	{113699, 321970, 33408}
{13, 18, 20}	{31795, 49602, 321970}
{13, 20, 40}	{31795, 321970, 197637}

Tab. 9.3.1-1. Grupy genów wynikające z reduktów o długości 6

Wzorce wyskakujące mogą być wykorzystane do budowy profili ekspresji genów. W naszym przypadku wskazują, jakie poziomy ekspresji mają określone geny dla komórek zrakowaciących (baza danych D_2). Na zielono zostały zaznaczone, te redukty i wzorce, które wskazują te same grupy genów. Zaznaczone wzorce wyskakujące wynikają na podstawie twierdzenia 8.2-1 z istnienia odpowiednich reduktów. Ponieważ wzorce te mają wsparcie 100%, więc na przykładzie reduktów obserwujemy tezę twierdzenia 8.2-3. Mówi ona o istnieniu zbiorów rozróżniających równych zbiorom atrybutów związanych z tymi wzorcami.

Wzorzec wyskakujący zgodnie z numeracją genów	Wzorzec wyskakujący zgodnie z nazwami genów
{(8,1),(10,1),(20,0)}	{(42837,1),(113699,1),(321970,0)}
{(8,1),(15,1),(20,0),(40,1)}	{(42837,1),(110608,1),(321970,0),(197637,1)}
{(8,1),(20,0),(30,1),(40,1)}	{(42837,1),(321970,0),(115012,1),(197637,1)}
{(8,1),(20,0),(39,1),(40,1)}	{(42837,1),(321970,0),(140870,1),(197637,1)}
{(10,1),(13,1),(20,0)}	{(113699,1),(31795,1),(321970,0)}
{(10,1),(20,0),(22,1)}	{(113699,1),(321970,0),(33408,1)}
{(13,1),(18,1),(20,0)}	{(31795,1),(49602,1),(321970,0)}
{(13,1),(20,0),(40,1)}	{(31795,1),(321970,0),(197637,1)}
{(15,1),(20,0),(22,1),(40,1)}	{(110608,1),(321970,0),(33408,1),(197637,1)}
{(20,0),(22,1),(30,1),(40,1)}	{(321970,0),(33408,1),(115012,1),(197637,1)}
{(20,0),(22,1),(39,1),(40,1)}	{(321970,0),(33408,1),(140870,1),(197637,1)}
{(5,1),(13,1),(20,0)}	{(47345,1),(31795,1),(321970,0)}

Tab. 9.3.1-2. Profile ekspresji genów wynikające z minimalnych wzorców wyskakujących o wsparciu 100% (0 – niski poziom, 1 – wysoki poziom)

W temacie znajdowania grup genów współzależnych nie dysponujemy badaniami innych grup badawczych dla tego zbioru.

Wybrane reduktory o długości 6 dla zbioru „Princeton” zostały zebrane w tabeli 9.3.1-3. Wybór został dokonany arbitralnie. Przedstawienie ma celu pokazanie postaci wzorców i omówienie wynikających z niej problemów.

Redukt zgodnie z numeracją genów	Redukt zgodnie z nazwami genów
{1, 2, 4, 10, 19, 26}	{M63391, M76378, J02854, H08393, M36634, T47377}
{1, 2, 10, 15, 19, 26}	{M63391, M76378, H08393, H40095, M36634, T47377}
{1, 2, 15, 19, 22, 26}	{M63391, M76378, H40095, M36634, T57619, T47377}
{1, 3, 4, 10, 19, 26}	{M63391, R87126, J02854, H08393, M36634, T47377}
{1, 3, 7, 19, 26, 27}	{M63391, R87126, Z50753, M36634, T47377, R84411}
{1, 3, 10, 13, 19, 26}	{M63391, R87126, H08393, J05032, M36634, T47377}
{1, 3, 10, 14, 19, 26}	{M63391, R87126, H08393, X63629, M36634, T47377}
{1, 3, 10, 15, 19, 26}	{M63391, R87126, H08393, H40095, M36634, T47377}
{1, 3, 10, 19, 26, 27}	{M63391, R87126, H08393, M36634, T47377, R84411}
{1, 3, 10, 19, 26, 29}	{M63391, R87126, H08393, M36634, T47377, X14958}
{1, 3, 10, 19, 26, 30}	{M63391, R87126, H08393, M36634, T47377, T62947}
{1, 3, 10, 19, 26, 31}	{M63391, R87126, H08393, M36634, T47377, H55916}
{1, 3, 19, 22, 26, 27}	{M63391, R87126, H08393, T57619, T47377, R84411}
{1, 3, 19, 22, 26, 30}	{M63391, R87126, H08393, T57619, T47377, T62947}
{1, 3, 19, 22, 26, 31}	{M63391, R87126, H08393, T57619, T47377, H55916}

Tab. 9.3.1-3. Grupy genów wynikające z reduktów o długości 6

W tabeli 9.3.1-4 zostały wyszczególnione profile wynikające z wzorców wyskakujących o wsparciu 81%.

Wzorzec wyskakujący zgodnie z numeracją genów	Wzorzec wyskakujący zgodnie z nazwami genów
{(15,0),(19,1),(26,0)}	{(H40095,0),(M36634,1),(T47377,0)}
{(3,1),(26,0),(27,0)}	{(R87126,1),(T47377,0),(R84411,0)}
{(3,1),(26,0),(30,0)}	{(R87126,1),(T47377,0),(T62947,0)}
{(3,1),(26,0),(31,0)}	{(R87126,1),(T47377,0),(H55916,0)}

Tab. 9.3.1-2. Profile ekspresji genów wynikające z minimalnych wzorców wyskakujących o wsparciu 81% (0 – niski poziom, 1 – wysoki poziom)

W tabeli 9.3.1-1 zostały zaznaczone te reduktory, które są nadzbiorami zbiorów atrybutów związanych z minimalnymi wzorcami wyskakującymi. Jak widać, każdy każdy wzorzec wynika z istnienia przynajmniej jednego z wyszczególnionych reduktów.

Autorzy w korekcie [29] zamieszczają listę 20 wzorców o największym wsparciu odpowiednio w klasie tkanek zdrowych i zrakowaciałych. W analizie rozważamy klasę komórek normalnych. Spośród podanych przez autorów wzorców wszystkie związane są przynajmniej z jednym genem, który nie pochodzi z listy zamieszczonej w tabeli 9.2.2-11. Geny te zostały najwyżej ocenione przez autorów. Ze względu na różne zbiory genów znaczących porównanie wzorców otrzymanych przez autorów ze znalezionymi w naszej analizie nie posiada wyraźnego uzasadnienia.

Profile uzyskane za pomocą wzorców wyskakujących są z powodzeniem wykorzystane w klasyfikacji tkanek. Jeśli chodzi o znajdowanie grup genów współzależnych zastosowanie nie jest tak jasne. Po pierwsze wyznaczone reduktory i wzorce poprzedza etap wyboru genów i dyskretyzacji, które prowadzą do straty dużej porcji informacji. Biorąc pod uwagę duży błąd towarzyszący stosowanej technologii może się okazać, że otrzymywane reduktory i wzorce stanowią szum. W szczególności wiele znalezionych zbiorów o małej liczbie atrybutów może mylnie sugerować istnienie zależności międzygenowych. Wiele atrybutów o dużej sile rozróżniającej może tworzyć redukt z różnymi grupami atrybutów.

Drużga kwestia dotyczy samej metody badania powiązań. Geny odpowiedzialne za realizację badanego procesu mogą ulegać ekspresji w różnych jego fazach. To sugerowałoby badanie kolejnych szeregów czasowych poziomów ekspresji oddzielnie dla każdej klasy tkanek.

9.3.3. Zbiór reduktów i przestrzeń wzorców wyskakujących

Badania dotyczyły porównania zbioru reduktów i lewego krańca płotu przestrzeni wzorców wyskakujących.

W przypadku zbioru „Ontario” występuje stosunkowo wysoka korelacja rozkładu wartości obu funkcji oceny genów (patrz rysunek 9.2.1-2). Współczynnik korelacji wynosi 0,716. Sugeruje to, że w przybliżeniu te same geny były istotne dla budowania wielu reduktów, co minimalnych wzorców wyskakujących. W szczególności obie funkcje najwyżej oceniły ten sam atrybut. Obserwujemy również korelację rozkładów ilości zbiorów określonej długości w

obu rozważanych rodzinach (patrz rysunek 9.2.1-3). Mogłoby to wskazywać na zbliżoną strukturę obu rodzin oraz pewne zależności pomiędzy badanymi pojęciami.

Dla zbioru „Princeton” nie obserwujemy już takich wyraźnych zależności. Korelacja pomiędzy rozkładami wartości funkcji oceny jest niższa (patrz rysunek 9.2.2-2). Współczynnik korelacji wynosi 0,631. Podobnie rozkłady częstotliwości występowania zbiorów o określonych długościach wyznaczone dla obu podejść nie pokrywa się ze sobą.

Rozbieżność w wynikach można tłumaczyć małą liczbą przeprowadzonych prób. Ponadto należy mieć świadomość, że w przypadku zbioru „Ontario” zbiór reduktów został wyznaczony w sposób przybliżony, co może mieć wpływ na obliczone statystyki. Dla zbioru „Princeton” obie rodziny zostały wyznaczone w sposób dokładny. Zwróćmy uwagę także na to, że przestrzeń wzorców wyskakujących zbioru „Ontario” zawiera wzorce o wsparciu 100%. Zgodnie z twierdzeniem 8.2-3 oznacza to, że istnieją zbiory rozróżniające złożone z atrybutów związanych z każdym z nich. Ponieważ wzorce pochodzą z lewego krańca płotu, więc uzyskane zbiory rozróżniające są jednocześnie reduktami. Fakt ten tłumaczy podobieństwo rozkładów wartości funkcji $attrval_{RED}$ i $attrval_{JEP}$.

Podsumowując, badania nie wykazały szczególnej zależności pomiędzy pojęciami reduktu i wzorca wyskakującego. Jedną z przyczyn stanowią ich przydatność do reprezentowania wiedzy. W szczególności zauważmy, że oba podejścia są silnie zależne od zbioru obiektów. Jeśli do uniwersum dodamy nowe elementy, to może to wpłynąć na zmianę charakteru rodzin. Część reduktów może przestać być reduktami, a część wzorców wyskakujących stanie się wzorcami wyłaniającymi o wysoki współczynniku wzrostu wsparcia. Szczególnie w stosunku do reduktów pojawia się wiele zarzutów, co do ich wrażliwości na rozszerzanie zbioru obiektów. Jeśli chodzi o wzorce wyskakujące dysponujemy bardzo dobrymi wynikami w zadaniach klasyfikacji, które sugerują, że są one znacznie bardziej odporne na tę operację. Z tego względu warto się zastanowić nad porównaniem reduktów dynamicznych z wzorcami wyskakującymi lub wzorcami wyłaniającymi o dużym współczynniku wzrostu wsparcia.

10. Opis programu testowego Microarray Analyser

10.1. Informacje ogólne

Aplikacja Microarray Analyser została przygotowana dla wygodnej analizy danych zgodnie ze ścieżką zaproponowaną w pracy. Ma ona za zadanie ułatwić testowanie zbioru danych pochodzących z eksperymentu z wykorzystaniem mikromacierzy za pomocą poruszonych metod. Nie stanowi środowiska do wstępnej obróbki pomiarów, gdyż posiada silne założenia na temat charakteru zbioru wejściowego.

Program jest wyposażony w interfejs graficzny, przy czym wszystkie wyniki cząstkowe są dokumentowane w logu. W ten sposób można wygodnie obserwować proces analizy zbioru danych.

Na program składają się pakiety:

- ep – implementacja algorytmu wyznaczania wzorców wyskakujących
- genetic – biblioteka algorytmów ewolucyjnych, napisana na potrzeby pracy, wzorowana na JGAP
- jrs – implementacja algorytmów wyznaczania reduktów
- microarray – implementacja aplikacji

Aplikacja została zrealizowana na platformie językowej JAVA. Do kompilacji kodu wykorzystano zestaw J2SE 1.4.2, a do wykonania użyto maszyny wirtualnej JVM 1.4.2. Oba są produktami firmy Sun Microsystems. Ze względu na charakter języka aplikacja jest niezależna od platformy systemowej i może być uruchomiona na każdej platformie, dla której dostępna jest odpowiednia maszyna wirtualna. Ponieważ pliki logu tworzone są w lokalnym katalogu, więc uruchomienie programu wymaga przegrania na dysk z prawem do zapisu.

10.2. Format danych wejściowych

Eksperymenty są prowadzone i dokumentowane w różny sposób. Wybrany zbiór często wymaga dodatkowego przygotowania. Zbiór wejściowy programu musi mieć następujące cechy:

- poziomy ekspresji są znormalizowane pomiędzy kolejnymi eksperymentami
- poziomy genów kontrolnych zostały usunięte
- wszystkie wartości są kompletne

Pojedynczy zbiór wejściowy zapisany jest w 5 plikach. Pliki zawierają opis tablicy decyzyjnej oraz nazwy genów i klas decyzyjnych. Tablica zapisana jest w przyjętej powszechnie konwencji, tzn. geny są wierszami a przykłady kolumnami. Przy czym w naszej analizie wiersze odpowiadają atrybutom a przykłady – obiektom tablicy. Zarówno geny, jak i przykłady są jednoznacznie uporządkowane. Porządek ten wykorzystywany jest przy oddzielnym opisie poziomów ekspresji, klas decyzyjnych przykładów i nazw genów. Ze względu na charakter analizy przykłady nie są opisywane przez żadne dodatkowe informacje poza swoją klasą decyzyjną.

Kolejnym możliwym klasom decyzyjnym są przypisywane kolejne liczby ze zbioru $N \cup \{0\}$. Przyporządkowanie to wykorzystywane jest do opisu klas przykładów.

Parametry opisujące rozmiar zbioru danych wykorzystywane przy opisie formatu plików:

- n – liczba genów
- m – liczba przykładów
- p – liczba klas decyzyjnych

Plik konfiguracyjny

Plik ma za zadanie wskazać wszystkie pliki opisujące eksperyment. Ponieważ pliki tworzą jedną całość, więc nie ma sensu oddzielnego wskazywania wszystkich plików przy każdorazowej analizie. Plik jest w formacie XML o strukturze przedstawionej w tabeli 11.2-1, wcięcia w kolumnie „znacznik” oddają hierarchię pojęć dokumentu.

Znacznik	Semantyka znacznika
Config	
matrixFile	Ścieżka pliku z macierzą ekspresji
classesFile	Ścieżka pliku z listą klas przykładów
geneNamesFile	Ścieżka pliku z nazwami genów
classNamesFile	Ścieżka pliku z nazwami klas

Tab. 11.2-1. Format pliku konfiguracyjnego zbioru danych

Plik ma rozszerzenie *.maa*.

Plik z macierzą ekspresji

Plik zawiera poziomy ekspresji wszystkich genów dla wszystkich przypadków. Plik jest w formacie tekstowym. Poziomy są liczbami rzeczywistymi w notacji z kropką. Kolejne wartości dotyczące ekspresji jednego genu są umieszczone w jednej linii i rozdzielone spacją. Każdy gen opisany jest w oddzielnej linii.

Plik ten jest zdecydowanie największym spośród zestawu plików opisujących eksperyment. Dokumentacja dostępna na stronach często zawiera pliki o zbliżonej postaci, co ułatwia proces przygotowania odpowiedniego formatu.

```

E11 E12 ... E1m
E21 E22 ... E2m
.....
.....
En1 En2 ... Enm

```

Rys. 11.2-1. Format pliku z macierzą ekspresji

Opis symboli formatu:

- Eij – poziom ekspresji i-tego genu w j-tym eksperymencie, liczba rzeczywista

Plik ma rozszerzenie *.matrix*.

Plik z listą klas przykładów

Plik zawiera wartości atrybutu decyzyjnego dla kolejnych przykładów. Klasa danego przykładu jest reprezentowana przez liczbę przypisaną tej klasie. Plik jest w formacie tekstowym. Wartości odpowiadające klasom kolejnych przykładów zapisane są w kolejnych liniach pliku.

```
klasa 1
klasa 2
.....
.....
klasa m
```

Rys 11.2-2. Format pliku z listą klas przykładów

Opis symboli formatu:

- klasa j – klasa decyzyjna j-tego przykładu, liczba z $\mathbb{N} \cup \{0\}$

Plik ma rozszerzenie *.classes*.

Plik z nazwami genów

Plik zawiera nazwy kolejnych genów. Plik ma charakter tekstowy. Nazwy są umieszczone w kolejnych liniach pliku.

```
nazwa genu 1
nazwa genu 2
.....
.....
nazwa genu n
```

Rys 11.2-3. Format pliku z nazwami genów

Opis symboli formatu:

- nazwa genu i – nazwa i-tego genu, łańcuch znaków

Plik ma rozszerzenie *.geneNames*.

Plik z nazwami klas

Plik zawiera nazwy kolejnych klas. Nazwy są uporządkowane zgodnie z relacją przyporządkowanym im liczb. Plik ma charakter tekstowy. Nazwy są umieszczone w kolejnych liniach pliku.

```
klasa 1
klasa 2
.....
.....
klasa p
```

Rys. 11.2-4. Format pliku z nazwami klas

Opis symboli formatu:

- nazwa klasy k – nazwa k-tej klasy, łańcuch znaków

Plik ma rozszerzenie *.classNames*.

10.3. Prowadzenie analizy zbioru danych

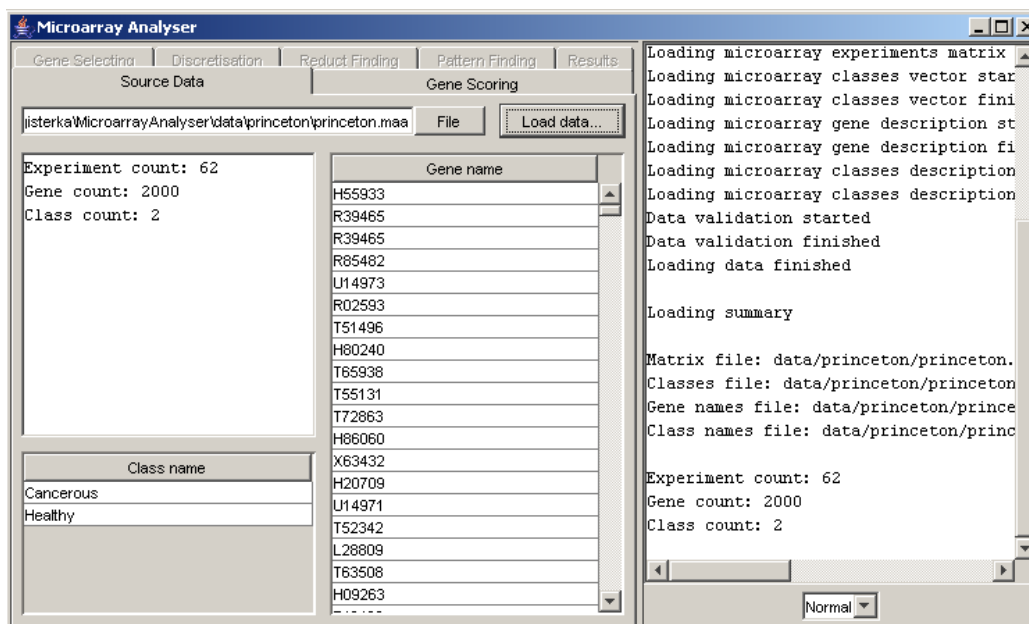
Struktura interfejsu graficznego odpowiada kolejnym krokom analizy. Możliwe jest wygodne ustawianie parametrów algorytmów realizujących poszczególne etapy oraz obserwacja rezultatów cząstkowych. Program pozwala na wielokrotne wywoływanie różnych kroków i rozpoczynanie w ten sposób analizy w nowym kierunku poczynając od pewnego punktu ścieżki.

Aplikacja jest typu SDI. Główne okno jest podzielone pionowo na dwie zasadnicze części. Prawa część zawiera okno wyświetlające log tworzony na bieżąco podczas sesji z programem. Zapisywane są w nim informacje na temat kolejnych kroków analizy. Możliwa jest także zmiana w czasie sesji poziomu szczegółowości informacji zapisywanych w logu. W ten sposób można dobrać sposób dokumentowania do celu badania zbioru danych.

Lewa część obejmuje system zakładek. Kolejne panele odzwierciedlają kroki prowadzonej analizy. Są one udostępniane sekwencyjnie. Panel jest aktywny dopiero wtedy, gdy zostały ukończone etapy dostarczające odpowiednio przetworzone dane. Z tego względu sposób prowadzenia analizy zostanie przedstawiony jednocześnie z omówieniem zakładek. Na niektórych zakładkach dostępnych jest kilka możliwych algorytmów do wyboru. Jeśli dany algorytm wymaga podania parametrów wejściowych, to objaśnienie ich znaczenia zostało umieszczone poniżej, przy opisie każdej z opcji. W programie automatycznie są podpowiadane wartości domyślne. Dla algorytmów zostały też wskazane rozdziały zawierające informacje na temat sposobu ich realizacji.

Wczytywanie danych

Rozpoczęcie analizy wymaga wskazania odpowiedniego pliku konfiguracyjnego danych wejściowych. W czasie wczytywania sprawdzana jest poprawność plików z właściwymi danymi. Proces sprowadza się do zainicjowania tablicy decyzyjnej o ciągłych atrybutach warunkowych oraz struktur mapujących nazwy genów i klas na reprezentacje wewnętrzną. Po przygotowaniu struktur informacje na ekran wyprowadzane są ogólne informacje na temat rozmiaru zbioru danych, badanych genów oraz możliwych klas decyzyjnych.



Rys. 11.3-2. Panel wczytywania danych wejściowych

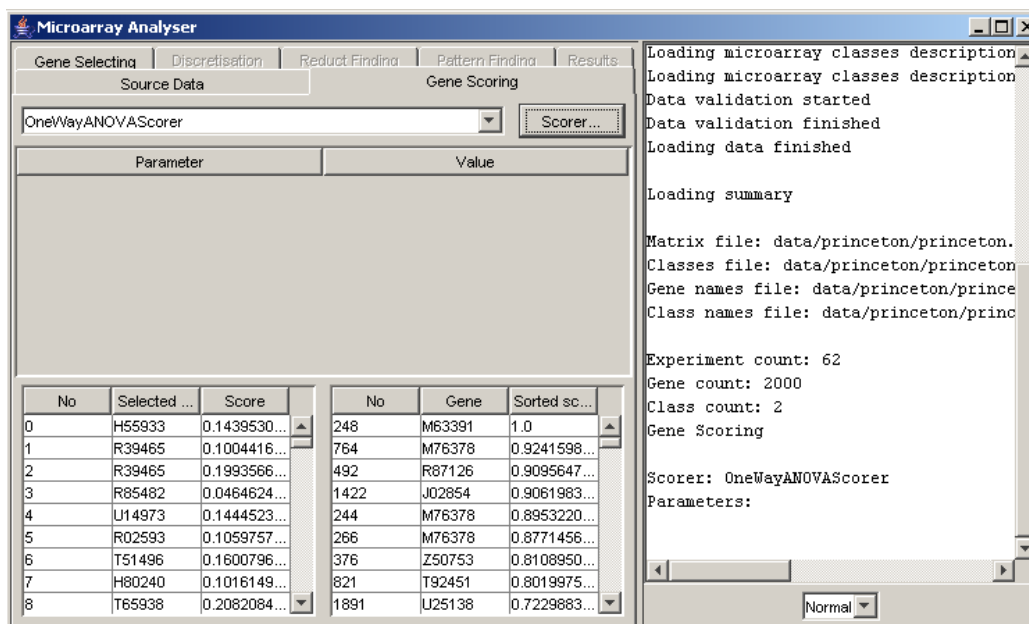
Ocena genów

Etap ten odpowiada za wyznaczenie ocen pojedynczych genów za pomocą jednej z wybranych metod. Dostępne są dwie metody oceny:

- NullScorer – trywialne przypisanie wszystkim atrybutom tej samej wagi równej 1
- OneWayANOVAScorer – metoda analizy wariancji wyliczająca wartość funkcji attrval_F dla każdego genu (rozdział 4.3.2)

Obie metody są bezparametrowe. Okno parametrów zostało pozostawione ze względu na ewentualne dodanie innych operatorów oceny atrybutów ciągłych.

Po uruchomieniu wybranej z metod wyznaczana jest funkcja oceny dla zbioru wszystkich genów. Wyniki zamieszczone są w dwóch tabelach. W jednej uporządkowane zgodnie z kolejnością początkową nazw genów, w drugiej zgodnie z nierosnącymi ocenami znaczenia.



Rys. 11.3-3. Panel oceny genów

Wybór zbioru znaczącego

Krok dotyczy wyboru genów, które będą rozważane w dalszej analizie. Zgodnie z ustawieniami użytkownika i na podstawie wyliczonych ocen genów wyznaczany jest zbiór znaczący. Na tej podstawie tworzona jest nowa tablica decyzyjna o zbiorze atrybutów warunkowych równym temu zbiorowi. Otrzymaną tablicę będziemy dalej nazywać zredukowaną tablicą decyzyjną.

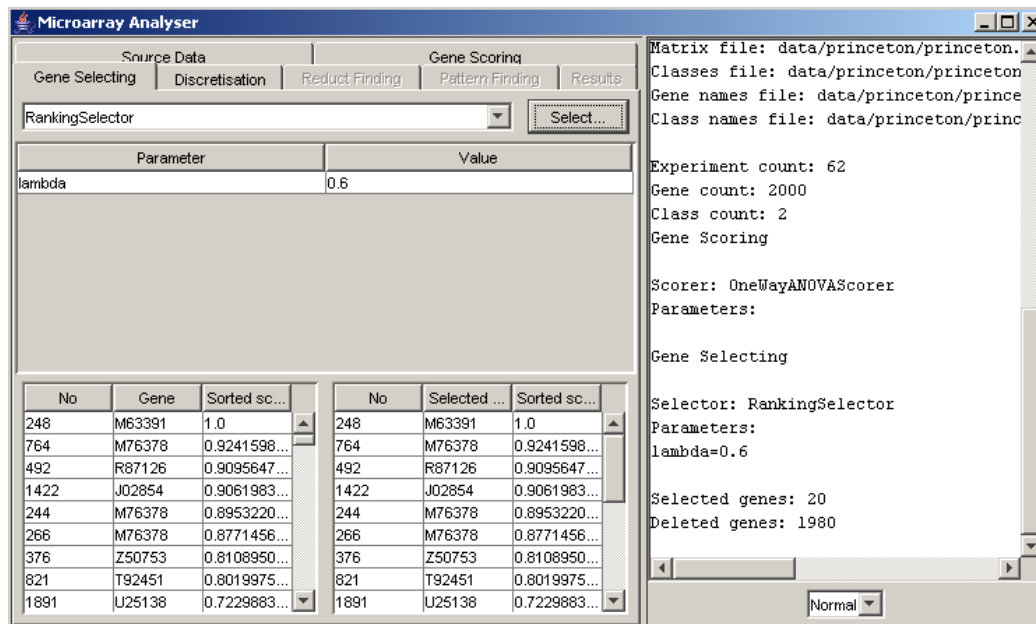
Wybór zbioru znaczącego może być dokonany za pomocą jednej z dwóch metod:

- RankingSelector – wyznaczenie zbioru λ , 1-rankingowo znaczącego (rozdział 4.4.3)
- SummingSelector – wyznaczenie zbioru λ , 1-sumacyjnie znaczącego (rozdział 4.4.3)

Obie metody wymagają podania rzeczywistego parametru λ z przedziału $\langle 0, 1 \rangle$.

W lewej tabeli zapisane są wszystkie geny pierwotnej tablicy decyzyjnej uporządkowane według nierosnących ocen znaczenia. Po wywołaniu algorytmu atrybuty zbioru znaczącego zostaną wypisane w prawej tabeli. Są one również uporządkowane według nierosnących ocen znaczenia.

Kryterium doboru zbioru jest jego rozmiar i rozkład ocen pojedynczych genów. Wyboru można dokonywać wielokrotnie dla różnych wartości parametru λ i w prawej tabeli obserwować postać zbioru znaczącego.



Rys. 11.3-4. Panel wyboru zbioru znaczącego

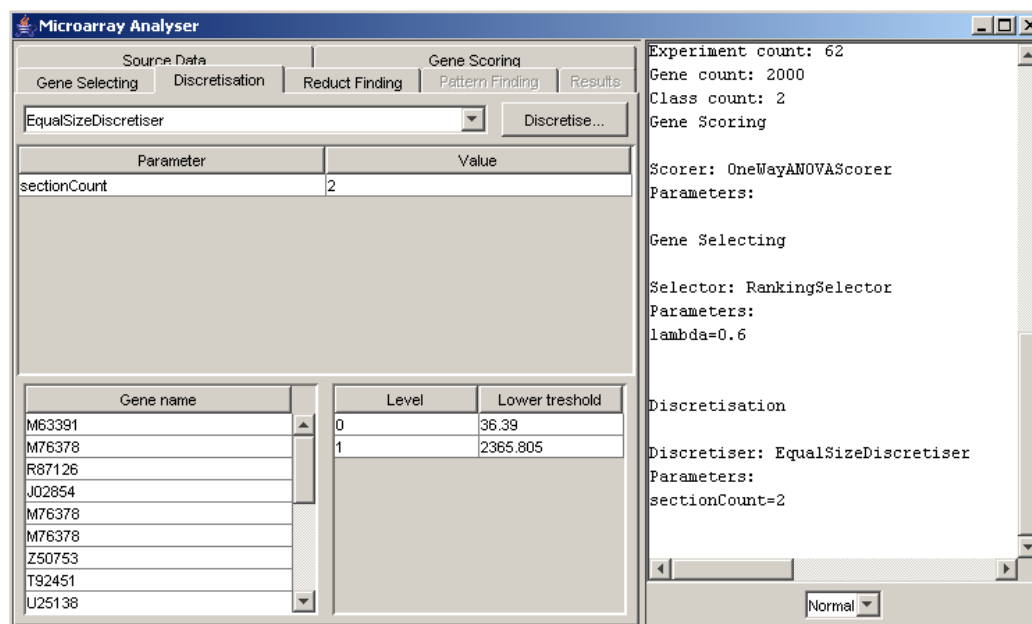
Dyskretyzacja

Etap jest niezbędny do zastosowania metod wymagających tablicy decyzyjnej o atrybutach dyskretnych. Zgodnie z obraną metodą zredukowana tablica decyzyjna jest dyskretyzowana do nowej tablicy, którą nazywać będziemy dalej zdyskretyzowaną tablicą decyzyjną. Przedziałom w podziale każdego atrybutu są przyporządkowywane niezależnie kolejne liczby ze zbioru $N \cup \{0\}$.

Dostępne są trzy metody dyskretyzacji:

- EqualSizeDiscretiser – dyskretyzacja z zachowaniem przedziałów tej samej szerokości (rozdział 5.3.1)
 - intervalCount –zamierzona liczba przedziałów w podziale dla każdego atrybutu, liczba całkowita
- EqualDepthDiscretiser – dyskretyzacja z zachowaniem przedziałów tej samej głębokości (rozdział 5.3.2)
 - intervalCount –zamierzona liczba przedziałów w podziale dla każdego atrybutu, liczba całkowita
- TopDownEntropyBasedDiscretiser – dyskretyzacja zstępująca analizująca entropię przy tworzeniu podziału dla każdego atrybutu (rozdział 5.3.3)

Lewa tabela zawiera nazwy genów, które odpowiadają atrybutom tablicy zredukowanej. Po dokonaniu procesu wskazanie myszą określonego genu w prawej tabeli wyświetlone zostaną przedziały wyznaczonego podziału zbioru wartości tego atrybutu.



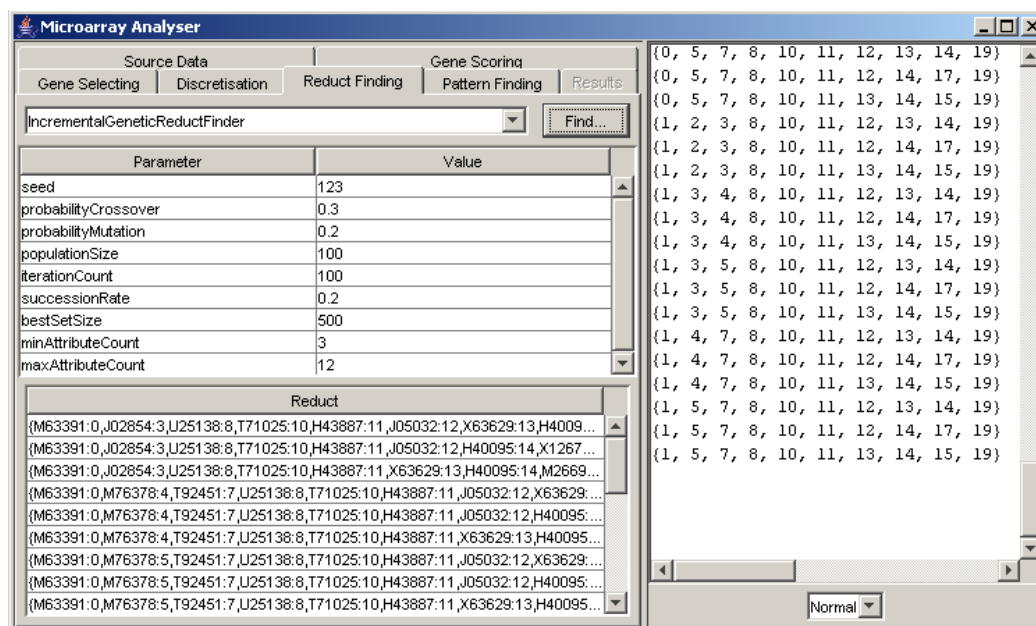
Rys. 11.3-5. Panel dyskretyzacji tablicy decyzyjnej

Wyznaczanie reduktów

W etapie tym wyznaczane są redukty zredukowanej tablicy decyzyjnej. Charakter znalezionych reduktów zależy od wyboru algorytmu ich znajdowania. Dostępne są dwa podejścia:

- ExactReductFinder – algorytm dokładny, wyznaczający wszystkie redukty tablicy (rozdział 6.4.1)
- IncrementalGeneticReductFinder – algorytm przybliżony wyznaczający zbiór reduktów o wartości nie większej niż zadana wartość (rozdział 6.4.4)
 - seed – oznaczająca ziarno dla generatora liczb losowych, liczba naturalna lub zero
 - crossoverRate – prawdopodobieństwo zajścia krosowania, liczba rzeczywista z przedziału $<0,1>$
 - mutationRate – prawdopodobieństwo zajścia mutacji, liczba rzeczywista z przedziału $<0,1>$
 - populationSize – rozmiar populacji, liczba naturalna
 - iterationCount – liczba generacji algorytmu ewolucyjnego, liczba rzeczywista z przedziału $<0,1>$
 - succesionRate – stopień zachodzenia procesu sukcesji, liczba rzeczywista z przedziału $<0,1>$
 - bestSetSize – rozmiar zbioru, w którym przechowywane są najlepsze osobniki, liczba naturalna
 - minAttributeCount – minimalna liczba atrybutów w przeszukiwanych zbiorach atrybutów
 - maxAttributeCount – maksymalna liczba atrybutów w przeszukiwanych zbiorach atrybutów, uzyskane redukty mogą mieć długość co najwyżej równą temu parametrowi

Po wykonaniu algorytmu wyznaczone redukty zostają umieszczone w tabeli.



Rys. 11.3-6. Panel wyznaczania reduktów

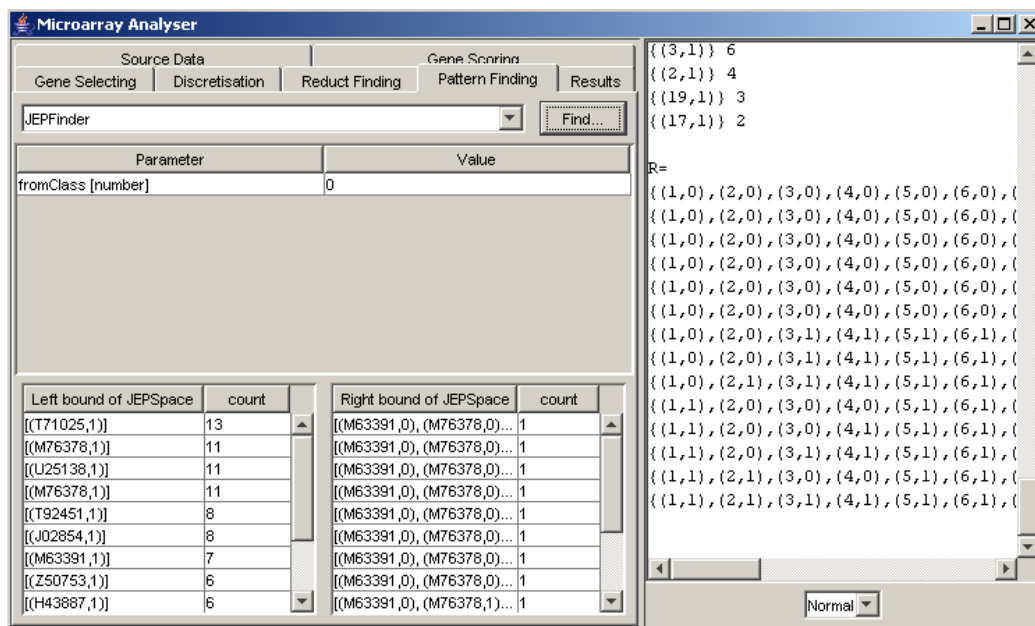
Wyznaczanie przestrzeni wzorców wyskakujących

W kroku tym wyznaczana jest cała przestrzeń wzorców wyskakujących dla rodziny baz danych dla zdyskretyzowanej tablicy decyzyjnej przy wyborze jednej klasy, jako klasy odniesienia. W pierwszym kroku budowana jest odpowiednia rodzina baz danych dla tablicy decyzyjnej. Następnie za pomocą algorytmu JEPProducer wyznaczana jest cała przestrzeń wzorców wyskakujących. Wzorce są wyznaczane z jednej wskazanej klasy do wszystkich innych, traktowanych jako jedna klasa. Przestrzeń zapisywana jest w reprezentacji płotowej. Algorytm ten wymaga podania parametru:

- fromClass – klasa, z której będą liczone wzorce

Po wykonaniu obliczeń zbiory przedmiotów z lewej krawca są zapisywane w lewej tabeli, a zbiory z prawego krawca – w prawej.

W interfejsie została przewidziana możliwość dodania innych algorytmów znajdowania wzorców wyskakujących, które na wyjściu dają zbiór wypukły.



Rys. 11.3-7. Panel wyznaczania przestrzeni wzorców wyskakujących

Raport wyników analizy

Ostatnia zakładka zawiera raport przeprowadzonej analizy. Po naciśnięciu przycisku generacji raportu zostają wyznaczone oceny pojedynczych genów ze zdyskretyzowanej tablicy decyzyjnej: attrval_X , $\text{attrval}_{\text{RED}}$, $\text{attrval}_{\text{JEP}}$. Razem z wyznaczoną na etapie eliminacji genów oceną dla attrval_F funkcje te są końcową informacją na temat znaczenia każdego z genów z osobna. Dla możliwych par rozkładów są obliczane współczynniki korelacji rozkładów. Wartości czterech funkcji oceny dla wszystkich genów zdyskretyzowanej tablicy są zamieszczone w tabeli. Informacje na temat porównania rozkładów są zamieszczone poniżej.

Microarray Analyser

Source Data | Gene Scoring

Gene Selecting | Discretisation | Reduct Finding | Pattern Finding | Results

No	Gene	In ANOVA	In chi	In reducts	In JEPs
0	M63391	1.0	0.48104569...	0.375	0.53846153...
1	M76378	0.92415989...	0.35410711...	0.625	0.30769230...
2	R87126	0.90956470...	0.44136741...	0.125	0.46153846...
3	J02854	0.90619830...	0.45995541...	0.5	0.61538461...
4	M76378	0.89532207...	0.52891286...	0.375	0.84615384...
5	M76378	0.87714568...	0.52891286...	0.375	0.84615384...
6	Z50753	0.81089501...	0.31788706...	0.0	0.46153846...
7	T92451	0.80199758...	0.45995541...	0.5	0.61538461...
8	U25138	0.72298835...	0.52891286...	1.0	0.84615384...
9	HNR393	n 70n41385	n 71964884	n n	n n

Size	Reducts	Patterns
1	0.0	0.3356643356643357
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
5	0.0	0.0
6	0.0	0.0
7	0.0	0.0
8	0.0	0.0
9	0.125	0.0
10	n 875	n n

Results

Normal

Rys. 11.3-8. Panel raportu wyników analizy

10.4. Propozycje rozwoju

Aplikacja została zaprojektowana z ideą łatwego dodawania alternatywnych operacji dla każdego z etapów. Zestaw interfejsów pozwala na wygodne zdefiniowanie nowych metod oceniania i wyboru genów, dyskretyzacji, wyznaczania reduktów i wzorców wyskakujących.

W klasie opisującej operator można wyspecyfikować:

- charakter jego parametrów (nazwy, typy i wartości domyślne)
- informacje zapisywane w logu

Struktura aplikacji pozwala także na naturalne dodanie kolejnych etapów analizy poprzez zdefiniowanie kolejnych zakładek w lewej części okna programu.

Proponowane rozszerzenia:

1. Dostosowanie wejścia aplikacji do różnych interfejsów. Przykładowo, umożliwienie bezpośredniego dostępu do najbardziej znanych baz mikromacierzy oraz obsługa ważniejszych formatów zapisu.
2. Wprowadzenie etapu uzupełnienia brakujących wartości. Implementacja różnych spotykanych w literaturze metod. Dodatkowo umożliwienie specjalnego traktowania brakujących wartości w całej analizie.
3. Wprowadzenie etapu normalizacji danych. Implementacja wybranych metod proponowanych dla analizy mikromacierzy. Związane jest to z rozszerzeniem informacji wejściowych o informacje na temat budowy macierzy, położenia komórek zawierających określone sekwencje, istnienia replikantów.
4. Analiza podzbiorów atrybutów przy wyznaczaniu zbioru znaczącego. Podejście wymaga wyboru sposobu oceny dowolnego podzbioru atrybutów warunkowych. W tym celu może być przykładowo wykorzystany odpowiedni klasyfikator (rozdział 4.4.1.1). Klasyfikator powinien operować na wartościach ciągłych, np. sieć neuronowa, maszyna wektorów nośnych.
5. Implementacja nowych metod dyskretyzacji. Przykładowo metod wstępujących lub metod opartych na badaniu błędu klasyfikacji. W ten sposób aplikacja może posłużyć celowi porównania zaproponowanych metod. Odniesieniem może być np. postać ostatecznych produktów analizy, czyli reduktów i wzorców wyskakujących, a także funkcje oceny pojedynczych genów.
6. Porównanie w stosunku do wyników publikowanych. Jeśli dostępne są wyniki otrzymane przez autorów, naturalnym rozszerzeniem byłoby możliwość odniesienia do nich rezultatów prowadzonej analizy. Biorąc pod uwagę niektóre z obecnie rozważanych tematów możliwe jest porównanie funkcji oceny pojedynczych genów i znalezionych wzorców wyskakujących.
7. Wizualizacja wyników w postaci wykresów punktowych i histogramów. Wykresy mogą dotyczyć korelacji ocen genów oraz przedstawienia statystycznych aspektów zbioru reduktów i przestrzeni wzorców wyskakujących.

8. Analiza grup genów. Wykorzystanie wyliczonych reduktów, wzorców wyskakujących oraz ocen statystycznych do wskazania jednej rodziny grup genów najsilniej związanych z badaną hipotezą.

11. Podsumowanie

Pierwszym zadaniem pracy było nakreślenie ram rozpatrywanej dziedziny. W tym celu wyłożone zostały biologiczne podstawy procesu ekspresji genów. Objęły one budowę i znaczenie informacji genetycznej zawartej w DNA w funkcjach życiowych organizmu. Bardziej szczegółowo omówiono etapy biosyntezy białek. Temat rozbudowano pod kątem najistotniejszych technik z zakresu inżynierii genetycznej. W szczególności poruszone zostało klonowanie sekwencji genetycznych oraz budowa bibliotek genowych. Wprowadzony aparat pojęciowy stał się punktem wyjścia do przybliżenia technologii badania ekspresji genów.

Wiele miejsca poświęcono opisowi technologii mikromacierzy. Przedstawiono najczęściej stosowane typy płytek oraz skojarzone z nimi warianty eksperymentów. Na schemacie doświadczenia pokazano ogólną ideę przyświecającą całej technologii. Szczegółowo zostały omówione etapy pomiaru poziomu ekspresji oraz normalizacji danych. Zarysowane podejście zilustrowano na przykładzie rzeczywistego eksperymentu dotyczącego badania nowotworów, który został wykonany na University of Toronto.

Obok informacji związanych bezpośrednio z praktyką prowadzenia badań poruszony został problem standaryzacji otrzymywanych wyników. W szczególności zarysowano temat gromadzenia informacji w postaci baz mikromacierzy oraz obecną postać standardu MIAME.

Po wprowadzeniu do dziedziny wskazane zostały najpopularniejsze kierunki analizy danych o ekspresji genów. W ten sposób pokazano tło dla prezentowanych w pracy rozważań. Proponowana ścieżka analizy dotyczy uczenia się pod nadzorem, przy czym wiedza na temat zależności genów jest wysnuwana przez badanie przestrzeni wektorów ekspresji odpowiadających kolejnym próbom eksperymentu. Zakładamy, że eksperyment został przeprowadzony poprawnie i że wszystkie dane na temat poziomów ekspresji są zapisane w postaci tablicy decyzyjnej. Wszelkie występujące w tekście definicje i twierdzenia zostały tak dobrane, aby tworzyć jednolitą całość. W szczególności pojęcie wzorców wyłaniających wymagało powiązania pojęcia bazy danych i zbioru przedmiotów z opisem danych, jaki proponuje tablica decyzyjna.

Problem oceny znaczenia genów odgrywa istotną rolę w przypadku badania ekspresji dużego zakresu genomu. Pozwala on na odfiltrowanie genów, które najprawdopodobniej nie są związane z rozważaną hipotezą oraz umożliwia wykorzystanie w dalszych etapach bardziej zaawansowanych, lecz słabo skalowalnych metod analizy.

Do oceny atrybutów ciągłych została wykorzystana jednoczynnikowa analiza wariancji. Jest to jedno z najpopularniejszych podejść w tym zakresie. Stanowi również punkt wyjścia dla proponowanych w literaturze bioinformatycznej modeli statystycznych o większej liczbie stopni swobody. Jako sposób wartościowania atrybutów dyskretnych przedstawiono statystykę χ^2 . Została ona wykorzystana do oceny genów po etapie dyskretyzacji.

Ocena genów daje możliwość wybrania odpowiedniego podzbioru do dalszej analizy. W wielu publikacjach stosuje się proste odcięcie genów o wartości nie mniejszej niż zadany próg. W pracy zostało zaproponowane ogólne podejście wartościowania podzbiorów atrybutów. Ma ono na celu wskazać nadzbiór wszystkich istotnych dla dalszej analizy zbiorów atrybutów. Metoda sprecyzowana została w pojęciach zbioru rankingowo i sumacyjnie znaczącego. Rozważono zarówno obliczenia dokładne, jak i oszacowanie przybliżone. W badaniach wykorzystane zostały jedynie przybliżenia wartości zbiorów oparte na ocenie należących do nich atrybutów. Ten szczególny przypadek jest zbliżony do

wspomnianego odcinania atrybutów według progu. Dla wyliczania oceny zbiorów zostały podane odpowiednie algorytmy.

Wiele algorytmów wymaga transformacji tablicy decyzyjnej z atrybutami ciągłymi do postaci zdyskretyzowanej. Etap ten prowadzi do straty dużej części informacji pochodzących z pomiaru. W praktyce wykorzystuje się najczęściej jedno z trzech podejść: dyskretyzacji z zachowaniem szerokości przedziałów, dyskretyzacji z zachowaniem głębokości przedziałów oraz dyskretyzacji zstępującej. Ostatnia metodą prowadzi proces pod nadzorem i wykorzystuje entropię do wyboru progów podziału. Zalecany warunek stopu dla tworzenia kolejnych przedziałów oparty jest o zasadę minimalnej długości kodu. Ze względu na uwzględnienie hipotezy docelowej podejście to powinno być preferowane.

Celem analizy zbioru danych jest wskazanie najistotniejszych genów oraz grup genów o współzależnej ekspresji. W pracy zostały wykorzystane dwa różne podejścia związane z analizą atrybutów dyskretnych: reduktów względne tablicy decyzyjnej oraz wzorce wyskakujące dla pary baz danych. Na tle tematyki biologicznej została podjęta próba porównania obu sposobów reprezentacji wiedzy.

W celu zarysowania problemu wyznaczania reduktów zostały podane podstawy teorii zbiorów przybliżonych. Problem ten został sprowadzony do problemu wyznaczania implikantów pierwszych pewnej funkcji boolowskiej oraz wyznaczania zbiorów trafieniowych dla pewnej rodziny zbiorów. Druga redukcja została potwierdzona formalnie. Dla problemu wyznaczania wszystkich reduktów tablicy decyzyjnej zostały przedstawione znane z literatury rozwiązania: dokładne, zachłanne oraz ewolucyjne. Zaproponowana została nowa metoda oparta o redukcję problemu do znajdowania zbiorów trafieniowych o minimalnej długości i dekompozycję przestrzeni poszukiwań. Dla rozwiązania podproblemów związanych z podprzestrzeniami wybrano metodę losową oraz ewolucyjną. Opisana metoda jest efektywniejsza w porównaniu z rozwiązaniami przeszukującymi całą przestrzeń, szczególnie w przypadku tablic decyzyjnych o dużej liczbie atrybutów. Ze względu na właściwość rozróżniania obiektów pomiędzy klasy reduktów są interpretowane jako zbiory genów współzależnych.

Drugie podejście oparte zostało o pojęcie wzorców wyłaniających. Jak zostało wcześniej zaznaczone, wszystkie podstawowe definicje i twierdzenia związane z tą formą wiedzy zostały sformułowane w odniesieniu do pojęcia tablicy decyzyjnej. Wprowadzane twierdzenia opatrzone dowodami. W pracy zajęliśmy się szczególnym przypadkiem wzorców wyłaniających – wzorcami wyskakującymi. W celu zwięzłej reprezentacji tego typu wzorców wykorzystano pojęcie płotu. Omówiono algorytm jep-producer, który w sposób dokładny wyznacza przestrzeń wzorców wyskakujących dla pary baz danych. Szczegółowo została omówiona podstawowa operacja algorytmu realizująca różnicę płotów o określonej postaci. Wzorce wyskakujące dobrze sprawdzają się w zadaniach klasyfikacji. Wzorce o wysokim wsparciu zostały wybrane jako grupy genów współzależnych.

Rezultaty otrzymane zgodnie z zastosowanymi podejściami wymagały wzajemnego porównania i odniesienia do wyników opublikowanych przez inne zespoły badawcze. W zakresie oceny pojedynczych genów zostały zaproponowane dwie funkcje oceny oparte o częstotliwość występowania genów w reduktach i w przestrzeni wzorców wyskakujących. Funkcje te zostały odniesione do ocen wynikających z analizy wariancji i statystyki χ^2 . Dla możliwych par rozkładów wartości tych funkcji zostały policzone korelacje.

Jako grupy genów współzależnych zgodnie z oboma podejściami zostały wybrane odpowiednio: najkrótsze redukty oraz wzorce należące do lewego krańca płotu przestrzeni wzorców wyskakujących. Dodatkowo cały zbiór reduktów oraz przestrzeń wzorców zostały porównane pod kątem liczności występowania w nich elementów o określonych rozmiarach. Wyniki zostały odniesione do opublikowanych wyników badań, które zakładały inne metody analizy tych samych danych.

W zakresie porównania pojęć reduktu i wzorca wyskakującego zostało formalnie pokazane, że pojedynczy redukt wyznacza zbiór wzorców wyskakujących oraz że wzorce wyskakujące o wsparciu 100% implikują cechę rozróżnialności dla zbiorów atrybutów z nimi związanych.

Wyniki eksperymentów nie wskazały na bliski związek pojęcia reduktu i wzorca wyskakującego. Zarówno zdefiniowane oceny atrybutów, jak i częstotliwość występowania zbiorów określonych rozmiarów w zbiorze reduktów i przestrzeni wzorców wyskakujących były ze sobą odpowiednio słabo skorelowane. Otrzymane w obu podejściach grupy genów różnią się od siebie i wyników dostępnych w publikacjach.

Należy wyraźnie podkreślić, że wysnuwanie wniosków na podstawie analizy wybraną metodą powinno być czynione bardzo ostrożnie. Po pierwsze dwie porównywane ścieżki mogą używać innej metody znajdowania genów znaczących, co może prowadzić do rozważania w następnych etapach innych zbiorów znaczących. Po drugie dane pochodzące z eksperymentów wykorzystujących mikromacierze są obarczone dużym błędem. Niesiona informacja możemy być dodatkowo zaciemniona poprzez proces dyskretyzacji. W takim przypadku otrzymane dalej rezultaty mogą okazać się fałszywe. Przykładem mogą tu być liczne krótkie redukty.

Ponadto, jest wysoce prawdopodobne, że geny odpowiadające wspólnie za dany proces uaktywniają się w jego kolejnych etapach. W takiej sytuacji nie można wykryć powiązań za pomocą analizy genów w wybranej chwili. Uzasadnione w tym wypadku wydaje się badanie szeregów czasowych ekspresji genów danego organizmu. Do analizy wzorców temporalnych może zostać wykorzystane zarówno podejście z zakresu teorii zbiorów przybliżonych, jak i wzorców wyłaniających.

Przeprowadzona analiza pokazała, że stosowanie zaawansowanych metod odkrywania wiedzy powinno być poprzedzone dokładnym etapem normalizacji i analizy danych ciągłych. Niezbędne jest także wprowadzenie dodatkowych miar błędu systematycznego dla każdego poziomu ekspresji. Dla porównywalności wyników z różnych eksperymentów potrzebna jest standaryzacja przetwarzania wstępnego i oceny genów. W obecnej chwili analiza zdyskretyzowanych poziomów ekspresji genów może prowadzić do mało wiarygodnych wyników.

12. Materiały

12.1. Bibliografia

- [1] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. Proc Natl Acad Sci, 96:6745-6750, 1999
- [2] Andrews J, Bouffard GG, Cheadle C, Lu J, Becker KG, Oliver B. *Gene discovery using computational and microarray analysis of transcription in the Drosophila melanogaster testis*. Genome Res. 2000 Dec;10(12):1841-2.
- [3] Arabas Jarosław. *Wykłady z algorytmów ewolucyjnych*. Wydawnictwa Naukowo-Techniczne, 2001
- [4] Baldi Pierre, Long Anthony D. *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes*. Bioinformatic, Vol. 17, no. 6 (2001), p. 509-519
- [5] Banachowski L., Diks K., Rytter W. *Algorytmy i struktury danych*. Wydawnictwa Naukowo-Techniczne, 1996
- [6] Bayardo Jr., Roberto J. *Efficiently mining long patterns from databases*. ACM-SIGMOD International Conference on Management of Data, 85-93, 1998
- [7] Bazan Jan G. *Metody wnioskowań aproksymacyjnych dla syntezy algorytmów decyzyjnych*. Praca doktorska pod kierunkiem prof. dr hab. Andrzeja Skowronka.
- [8] Bjorvand Anders Torvill, Komorowski Jan. *Practical applications of genetic algorithms for efficient reduct Computation*. The Norwegian University of Science and Technology - Department of Computer and Information Science, 1997
- [9] Brazma Alvis, Vilo Jaak. *Gene expression data analysis*. FEBS Letters 480 (2000) 17-24, 2000
- [10] Cichosz Paweł. *Systemy uczące się*. Wydawnictwa Naukowo-Techniczne, 2000
- [11] Cormen Thomas, Leiserson Charles, Rivest Ronald. *Introduction to Algorithms*. The Massachusetts Institute of Technology, 1994
- [12] Cykier Adam. *Implikanty pierwsze funkcji boolowskich, metody wyznaczania i zastosowania*. Uniwersytet Warszawski, Instytut Matematyki, Informatyki i Mechaniki, 1997
- [13] Dong Guozhu, Li Jinyan. *Efficient mining of emerging patterns: discovering trends and differences*. ACM SIGKDD 1999 International Conference on Knowledge Discovery & Data Mining, pages 43 – 52, 1999
- [14] Dong Guozhu, Zhang Xiuzhen, Wong Limsoon, Li Jinyan. *CAEP: Classification by Aggregating Emerging Patterns*. Proceedings of Second International Conference on Discovery Science, Lecture Notes in Computer Science, Vol. 1721 Springer, 1999
- [15] Eckman BA, Aaronson JS, Borkowski JA, Bailey WJ, Elliston KO, Williamson AR, Blevins RA. *The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining*. Bioinformatics vol. 14 no. 1, 1998 (2-13)
- [16] Fayyad U., Irani K. *Multi-interval discretization of continuous-valued attributes for classification learning*. Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1022–1029, 1993.
- [17] Fleurent Charles, Ferland Jacques. *Genetic and hybrid algorithms for graph coloring*. Annals of Operations Research (331-362), 1992
- [18] Goldberg David, *Genetic Algorithms in search, optimization and machine learning*. Addison-Wesley Publishing Company, 1989

- [19] Gu Chi, Rao D., Stormo Gary, Hicks Chindo, Province Michael. *Role of gene expression microarray analysis in finding complex disease genes*. Genetic Epidemiology (23; 37-56), 2002
- [20] Han Jiawei, Kamber Micheline. *Data Mining: Concepts and Techniques*. Academic Press, 2001
- [21] Hui Shirley. Rough Set classification for gene expression data, University of Waterloo, 2002
- [22] Irizarry Rafael, Parmigiani Giovanni, Guo Mingzhou, Dracheva Tatiana, Jen Jin. *A statistical analysis of radiolabeled gene expression data*.
- [23] Jóźwiak Lech. *Information Relationships and Measures An Analysis Apparatus for Efficient Information System Synthesis*. 23rd EUROMICRO Conference '97 New Frontiers of Information Technology, 1997
- [24] Jóźwiak Lech. *Information relationships and measures in application to logic design*. Twenty Ninth IEEE International Symposium on Multiple-Valued Logic, 1999
- [25] Kerr Kathleen, Afshari Cynthia, Bennett Lee, Bushel Pierre, Martinez Jeanelle, Walker Nigel, Churchill Gary. *Statistical analysis of a gene expression microarray experiment with replication*. Nieopublikowany. Data pobrania zasobu: 15.04.2004.
Adres: <http://www.jax.org/staff/churchill/labsite/pubs/index.html>.
- [26] Kerr Kathleen, Churchill Gary. *Experimental design for gene expression microarrays*. Biostatistics, 2000
- [27] Li Jinyan. *Mining emerging patterns to construct accurate and efficient classifiers*. Praca doktorska, 2001
- [28] Li Jinyan, Dong Guozhu, Ramamohanarao Kotagiri. *Making Use of the Most Expressive Jumping Emerging Patterns for Classification*. Knowl. Inf. Syst. 3(2): 131-145 (2001)
- [29] Li Jinyan, Wong Limsoon. *Corrigendum. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns*. Bioinformatics vol. 18 no. 10, 2002 (1406-1407)
- [30] Li Jinyan, Wong Limsoon. *Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns*. Bioinformatics vol. 18 no. 5, 2002 (725-734)
- [31] Li Jinyan, Wong Limsoon. *Emerging patterns and gene expression data*. Genome Informatics 12: 3-13, 2001
- [32] Lin Li, Jiang Yunfei. *Computing minimal hitting sets with genetic algorithm*. Jinan, Sun Yat-sen University, Guangzhou, China
- [33] Lin Li, Jiang Yunfei. *Computing minimal hitting sets with genetic algorithm*. International Workshop on Principles of Diagnosis (DX-02), 2002
- [34] Lin Li, Jiang Yunfei. *The computation of hitting sets: review and new algorithm*. Information Processing Letters 86 (2003) 177-184
- [35] Michalewicz Zbigniew. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer Verlag, 1992
- [36] Momma Michinari, Song Minghu, Bi Jinbo. *Mining microarray gene expression data*. Rensselaer Polytechnic Institute, 2001
- [37] Mrózek Adam, Płonka Leszek. *Analiza danych metoda zbiorów przybliżonych. Zastosowania w ekonomii, medycynie i sterowaniu*. Akademicka Oficyna Wydawnicza PLJ, 1999
- [38] Murali T. M., Kasif Simon. *Extracting conserved gene expression motifs from gene expression data*. Pacific Symposium on Biocomputing 2003 (77-88)
- [39] Zdzisław Pawlak. *Rough Sets. Theoretical Aspects of Reasoning about Data*. Wydawnictwo Politechniki Warszawskiej, Warszawa, 1990

- [40] Quackenbush J. *Computational genetics computational analysis of microarray data*. Nat Rev Genet. 2001 Jun;2(6): 418-27
- [41] Rocke David, Durbin Blythe. *A Model for measurement error for gene expression arrays*. Journal of Computational Biology, Vol. 8, 557-569, 2001
- [42] Rymon R. *Search Through Systematic Set Enumeration*. International Conference on Principles of Knowledge Representation and Reasoning, 1992.
- [43] Singer Maxine, Berg Paul. *Dealing with Genes. The Language of Heredity*. University Science Books, 1992
- [44] Turner P.C. McLennan A.G., Bates A.D., White M.R.H. *Instant notes in Molecular Biology*. BIOS Scientific Publishers Limited, 1997
- [45] Walczak Zbigniew, Dominik Andrzej, Terlecki Paweł. *Space decomposition in the minimal reduct problem*. VII Krajowa Konferencja, Algorytmy ewolucyjne optymalizacja globalna, 2004
- [46] Wigle, D., Jurisica, I., N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, S. Keshavjee, G. Darling, T. Winton, B. Breitkreutz, P. Jorgenson, M. Tyers, F. A. Shepherd, M.S. Tsao. *Molecular profiling of non-small cell lung cancer and correlation with disease-free survival*. Cancer Research, 62(11):3005-3008, June 1, 2002.
- [47] Wilson Robin, *Introduction to Graph Theory*. Robin Wilson, 1996
- [48] Winter P.C., Hickey G.I., Fletcher H.L. *Instant notes in Genetics*. BIOS Scientific Publishers Limited, 1998
- [49] Wroblewski Jakub. *Finding minimal reducts using genetic algorithm*. University of Warsaw - Institute of Mathematics, 1995
- [50] Yang Yee, Dudoit Sandrine, Luu Percy, Lin David, Pend Vivian, Ngai John, Speed Terence. *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation*. Biostatistics, 2000

12.2. Zasoby internetowe

- [51] Oficjalna strona dotycząca Affymetrix. Data pobrania zasobu: 12.04.2004. Adres: <http://www.affymetrix.com/index.affx>
- [52] Oficjalna strona dotycząca Bioinformatic Sequence Markup Language (BSML). Data pobrania zasobu: 12.04.2004. Adres: <http://www.bsml.org>
- [53] Oficjalna strona dotycząca Kentucky Biomedical Research Infrastructure Network. Data pobrania zasobu: 12.04.2004. Adres: <http://www.kbrin.louisville.edu>
- [54] Oficjalna strona dotycząca Minimum Information About a Microarray Experiment (MIAME). Data pobrania zasobu: 12.04.2004. Adres: <http://www.mged.org/Workgroups/MIAME/miame.html>
- [55] Strona Igor Jurisica. University of Toronto Health Network, Ontario Cancer Institute. Data pobrania zasobu: 12.04.2004. Adres: <http://www.cs.toronto.edu/~juris>
- [56] Strona Nature Genetics. Data pobrania zasobu: 12.04.2004. Adres: <http://www.nature.com/genetics>
- [57] Strona Princeton University – gene expression project. Data pobrania zasobu: 12.04.2004. Adres: <http://microarray.princeton.edu/oncology/>
- [58] Strona UCI Machine Learning Repository. Data pobrania zasobu: 12.04.2004. Adres: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [59] Strona Y.F. Leung's. Functional Genomics. Data pobrania zasobu: 12.04.2004.

Adres: <http://ihome.cuhk.edu.hk/~b400559>