

XII Konferencja PLOUG
Zakopane
Październik 2006

Analiza narzędzia Data Mining ORACLE 10g do klasyfikacji komórek nowotworowych w cytometrycznym systemie skaningowym

Włodzimierz Stanisławski, Ewelina Szydłowska

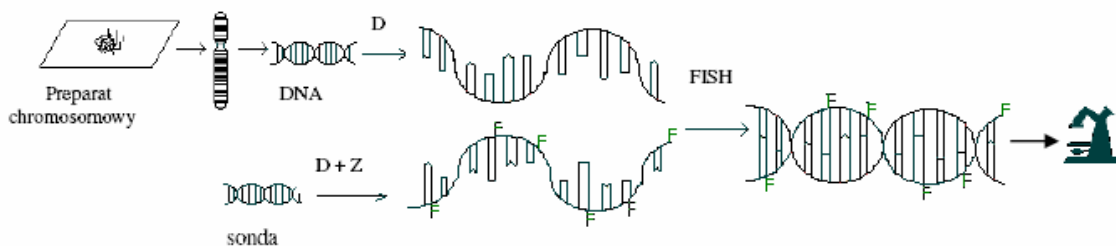
*Politechnika Opolska, Wydział Elektrotechniki, Automatyki i Informatyki
e-mail: stan@po.opole.pl, e.szydłowska@weia.po.opole.pl*

1. Wprowadzenie

Choroby nowotworowe znane są od wieków. Jednym z coraz częściej wykrywanych nowotworów jest rak pęcherza. Występuje on przeważnie u osób w wieku starszym (60 -70 lat) i stanowi czwarty co do częstości występowania nowotwór złośliwy u mężczyzn i ósmy u kobiet. Guzy pęcherza można podzielić na kilka grup. Najczęstszym typem raka jest rak z nabłonka przejściowego (ang. transitional cell carcinoma) stanowiący 90% wszystkich przypadków. Kolejne istotne grupy to rak płaskonabłonkowy i rak gruczołowy [BoSi] .

W leczeniu chorób nowotworowych stosuje się wiele metod. Zależą one od rodzaju guza i szybkości jego rozprzestrzeniania się. Powstawanie komórek nowotworowych jest procesem ciągłym. W trakcie jego trwania następuje przemiana komórki zdrowej w komórkę rakową. Nie wiadomo kiedy się rozpoczyna i jak długo trwa taki proces. W diagnostyce wykorzystywane są głównie metody inwazyjne, umożliwiające rozpoznanie choroby wówczas gdy jest widoczna makroskopowo [BCBK03] . Dlatego też istotne jest poszukiwanie rozwiązań dających możliwości jak najwcześniejszego wykrycia raka. Opracowane są różne metody badań jak: posiew moczu (laboratoryjne testy na obecność bakterii), cytologia moczu (mikroskopowe badania komórek wyeliminowanych z pęcherza), cytometria przepływowa (pomiar charakterystycznych fizycznych lub chemicznych cech komórek), cystoskopia (badanie pęcherza moczowego przy użyciu wziernika), biopsja (pobranie fragmentów tkanki do analizy komórek rakowych i identyfikacji typu nowotworu), urografia dożylna (wstrzykiwanie do krwiobiegu kontrastowego barwnika i przeprowadzenie zdjęcia rentgenowskiego) [AbboHt] . Duże nadzieje w tej dziedzinie łączy się z zastosowaniem biomarkerów choroby nowotworowej. Biomarkery to substancje produkowane przez nowotwory lub wytwarzane przez organizm w reakcji na obecność nowotworów w organizmie. Wykrywanie komórek rakowych z wykorzystaniem biomarkerów możliwe jest poprzez badanie krwi, moczu czy tkanki. Analizy takie mogą być powiązane z innymi badaniami jak cystoskopia, gdy monitorowanie pewnych części układu moczowego jest utrudnione bądź niemożliwe. Zastosowanie znajdują także we wczesnej diagnozie czy określaniu prawdopodobieństwa nawrotu choroby. [AbboCD]

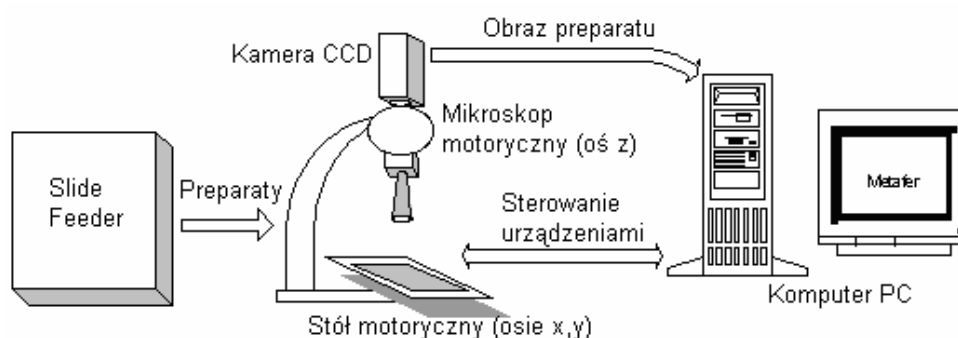
Jedną z metod wykorzystania markerów jest fluorescencyjna hybrydyzacja in situ (FISH). W metodzie tej markerem jest sekwencja DNA, którą uwidacznia się przez hybrydyzację z sondą fluorescencyjną. Do zastosowania tej metody wymagane jest rozdzielenie podwójnej helisy (denaturacja DNA, rys. 1) poprzez wysuszenie preparatu na szkiełku mikroskopowym i podziałanie na niego formamidem. W pierwszym etapie wprowadzania hybrydyzacji, sonda była znakowana promieniotwórczo. Ze względu na niewystarczającą skuteczność, a także względy zdrowotne i środowiskowe, pod koniec lat 80-tych wprowadzono fluorescencyjne znaczniki DNA. Zastosowanie znaczników o różnych kolorach emisji, umożliwiło hybrydyzację wielu sond z jednym chromosomem. Znaczniki fluorescencyjne o różnych kolorach (różne długości fal) włącza się do nukleotydów lub bezpośrednio do cząsteczki DNA. Do wykrywania wykorzystywane są najczęściej mikroskopy fluorescencyjne lub detektory fluorescencji. [Bro01]



Rys. 1. Schemat FISH (D- denaturacja, Z - znakowanie, F-fluorochrom) [ZaWi03]

2. Charakterystyka zbioru danych

Do przeprowadzenia badań stosuje się systemy komputerowe wyposażone w kamerę wideo i oprogramowanie do analizy cytogenetycznej. Przykładem takiego systemu jest „Metafer” firmy „MetaSystems”. Szczegółową budowę stanowiska laboratoryjnego oraz sposoby wykonywania pomiarów można znaleźć w literaturze [PILo01,Guz05]. Główne elementy stanowią mikroskop, kamera CCD oraz 8-mio pozycyjny stół skanujący. Analiza preparatów zaczyna się od ich podziału pod mikroskopem na obszary skanowania. Ostrość dobierana jest automatycznie. Pobrane za pomocą kamery obraz przesyłany jest do komputera, a następnie w wyniku analizy obrazu zostają wyodrębnione komórki, będące rezultatem skanowania (rys. 2).



Rys. 2. Komponenty systemu "Metafer" [Guz05]

Dostarczone oprogramowanie pozwala na analizę komórek pod względem wielu parametrów. Jednym z nich są morfometryczne cechy komórki. Opisują one takie właściwości jak rozmiar komórki, kształt, intensywność. Wybór parametrów zależy od rodzaju badanego materiału genetycznego. W oprogramowaniu zaimplementowane są narzędzia optymalizacji pozwalające na wybór najlepszych wartości tych parametrów. Wyniki analizy dostępne są w postaci plików tekstowych zawierających zbiór wartości wybranych parametrów. Pliki mogą być poddawane kolejnym etapom badań, których celem jest opracowanie procedur pozwalających na identyfikację komórki rakowej.

Do klasyfikacji komórek przygotowane są pliki treningowe. Zawierają one ocenione przez eksperta obrazy pól przebadanych preparatów. W każdym polu obiekty zostały wydzielone poprzez zaznaczenie komórek rakowych. W zbiorach danych dostarczonych do dalszej analizy, każdy obiekt opisany jest przy pomocy 217 atrybutów, na przykład:

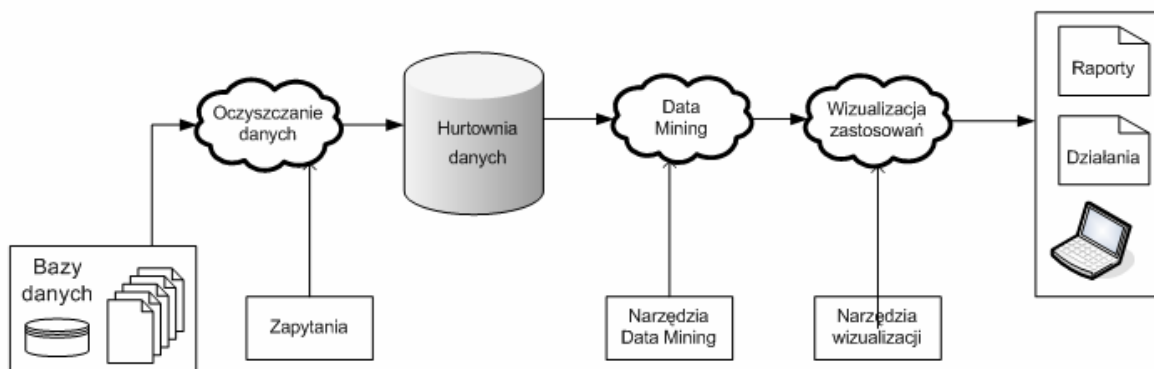
- numer obszaru z którego pochodzi dany obiekt
- informacja o typie obiektu (czy jest komórką rakową)
- minimalny i maksymalny obszar zawarty w konturze obiektu
- wklęsłość obiektu
- stosunek przekątnych obiektu

Do badań zostało przekazanych 16 zbiorów treningowych. Dalszą analizę można wykonywać na kilka sposobów. Pierwszy polega na wykorzystaniu jednego ze zbiorów jako zbioru uczącego, a następnie na testowaniu wygenerowanego modelu na pozostałych danych. Operacje takie można powtarzać wymieniając ze sobą zbiory treningowe. W innym podejściu, wykorzystanym w prezentowanej pracy, zbiory zostały ze sobą połączone. Ostateczny zbiór, składa się z 22962 obiektów. W zbiorze tym 640 obiektów zostało zidentyfikowanych przez eksperta jako komórki rakowe. Stanowi to niecałe 3% wszystkich dostępnych danych. Zbiór został losowo podzielony na trzy

części w proporcjach: 30%, 21%, 49%. Ze względu na tak liczne zbiory danych należy zastosować narzędzia, które pozwolą na identyfikację ważnych, oryginalnych, potencjalnie przydatnych i zrozumiałych wzorców.

3. Pojęcie eksploracji danych

Proces identyfikacji wzorców nazywamy odkrywaniem wiedzy (ang. Knowledge Discovery). Wzorec jest tutaj rozumiany w szerokim zakresie jako związki, korelacje, trendy, deskryptory rzadkich zdarzeń itp. Jedno z narzędzi, któremu w dalszej części zostanie poświęcona uwaga to eksploracja danych (ang. Data Mining). Pojęcia eksploracji danych i odkrywania wiedzy są czasami używane zamiennie [CPS00]. W prezentowanym artykule pojęcie eksploracji poruszane jest w odniesieniu do jednej z faz procesu odkrywania wiedzy (rys. 3).



Rys. 3. Proces odkrywania wiedzy [CPS00]

Dane gromadzone są w różnych postaciach baz. Mogą to być płaskie pliki, bazy relacyjne, obiektowe. Do procesu eksploracji trzeba je w odpowiedni sposób przygotować. Najlepiej gdy dane do procesu eksploracji pochodzą z hurtowni danych. Proces eksploracji danych gromadzi w sobie wiele narzędzi takich jak statystyka, modele regresji, sieci neuronowe, zbiory rozmyte, metody ewolucyjne, zbiory przybliżone, klastrowanie itp. Dane, które są wynikiem analizy można przedstawić za pomocą narzędzi wizualizacji.

Skupiając się tylko na procesie eksploracji, można wyróżnić w nim trzy istotne etapy:

- opisanie danych; na podstawie statystycznych analiz, wykresów można zaobserwować podstawowe właściwości danych
- zbudowanie i testowanie predykcyjnego modelu; na podstawie danych budowany jest model bazujący na poznanych wzorach danych. Tak zbudowany model jest testowany na innym zbiorze. Dobrze zbudowany model nie powinien się mylić, a jego wyniki muszą w dużym stopniu pokrywać się z rzeczywistymi wartościami
- doświadczalna weryfikacja modelu; gdy model zostanie już dokładnie zaprojektowany powinniśmy dokonać doświadczalnej weryfikacji, aby upewnić się czy można polegać na jego przewidywaniach.

Analiza danych nie jest zagadnieniem prostym. W trakcie tego procesu należy radzić sobie z różnymi problemami. Pierwszy z nich to ogromne rozmiary danych. Problem jest tutaj złożoność czasowa. Poszukiwanie zależności pomiędzy wartościami atrybutów wymaga stosowania metod heurystycznych lub zmniejszania obszaru poszukiwań. Zmniejszanie może odbywać się horyzontalnie lub w płaszczyźnie pionowej. Pierwszy przypadek polega na przeprowadzeniu dyskretyzacji

wartości cech, drugi polega na usuwaniu cech, które wydają się nadmierne. Kolejny problem z jakim należy się zmierzyć to dynamiczna natura danych. Zbiory danych są nieustannie modyfikowane poprzez dodawanie nowych elementów lub zamianę istniejących. Dlatego też narzędzia data mining powinny być cały czas rozwijane, a dostarczona wiedza powinna być przyrostowo aktualizowana. Istotnym kłopotem dla analizy są błędy niesystematyczne zwane szumami. Metody analizy nie powinny być zbyt czułe na tego typu zachowania, tak aby nie powodowały zakłócania reguł. Jeszcze innym problem w odkrywaniu wiedzy są brakujące wartości.

Wśród metod eksploracji danych można wyróżnić dwie typowe grupy zagadnień [Mor99, Mor05]. Biorąc pod uwagę wyniki analizy dostępne są techniki predykcyjne i deskrypcyjne. Pierwsze, na podstawie znalezionych wzorców dokonują przewidywań na przykład wartości atrybutów, zachowań czy cech. Znajdują więc one zastosowanie w przypadkach, gdzie istotne jest oszacowanie cechy wyjściowej. Techniki deskrypcyjne wykorzystują poznane wzorce do opisywania danych. Zawierają się tu przede wszystkim zagadnienia grupowania. Podziału eksploracji można także dokonać biorąc pod uwagę zbiór danych wejściowych. Wyróżniamy tutaj uczenie z nauczycielem i bez nauczyciela. W pierwszym przypadku dane wejściowe stanowią pewien zbiór uczący, gdzie dla określonego zestawu wartości atrybutów poznawane są wartości atrybutu wyjściowego. Drugi rodzaj uczenia, bez nauczyciela, występuje gdy nie posiadamy zbioru uczącego. Wtedy sformułowany jest model, najlepiej pasujący do obserwowanych danych.

4. Narzędzia eksploracji danych

Na rynku dostępnych jest coraz więcej narzędzi oferujących metody eksploracji danych. Dane mogą być przechowywane w różnych bazach i analizowane przez wiele narzędzi. Najkorzystniejsze są jednak takie rozwiązania, które łączą w sobie wiele funkcjonalności. Korzystanie z jednego narzędzia pozwala na obniżenie kosztów, a dostarczone informacje są bardziej rzetelne. W przeprowadzanych badaniach wykorzystywana jest baza danych Oracle w wersji 10g. Serwer ten posiada szereg funkcji umożliwiających rozbudowane analizy danych. Istotną cechą tego środowiska jest połączenie procesu odkrywania z systemem zarządzania bazą danych. Dzięki temu różne procesy jak przygotowywanie danych, ich transformacja, generowanie i wykorzystywanie modeli mogą się odbywać w jednym systemie bazy danych. Daje to także duże możliwości programistom, poprzez połączenie technik eksploracji z aplikacjami bazodanowymi. Komponentem oferującym analizy danych jest Oracle Data Mining (ODM). W jego skład wchodzi trzy elementy:

- Data Mining Engine (DME) – zapewnia infrastrukturę, zawierającą zestaw usług Data Mining udostępnianych dla klientów API
- interfejs aplikacji (API) – umożliwia dostęp do funkcji i algorytmów zaimplementowanych w DME
- repozytium metadanych – wykorzystywane poprzez DME do udostępniania obiektów wygenerowanych w trakcie analiz

W interfejsie aplikacji można wyodrębnić trzy części. Każda z nich skierowana jest do innego typu użytkownika. Pierwsza to Oracle Data Mining Predictive Analytics (PA). Jest to pakiet zawierający dwa programy: przewidywanie (ang. Predict), wyjaśnianie (ang. Explain). W przewidywaniu wykorzystywane są algorytmy klasyfikacji i regresji, a w wyjaśnianiu algorytm ważności atrybutów. Są one w pełni zautomatyzowane (użytkownik nie wybiera algorytmu, nie określa parametrów ustawień), wymagany jest tylko odpowiedni format danych wejściowych. Program przewidywania skierowany dla zwykłych użytkowników (nie technicznych), jak dyrektorzy marketingu, dla których głównym celem jest uzyskanie w krótkim czasie rzetelnych wyników. Drugi interfejs jest skierowany do programistów. Dostępny jest w dwóch językach: Java i PL/SQL. Pozwala on na wdrażanie wbudowanych algorytmów do aplikacji klienckich. Oba API są ze sobą kompatybilne (od wersji ODM 10.2), tak więc można budować modele na przykład z wykorzysta-

niem skryptów PL/SQL, a testować przy użyciu aplikacji Javy. Dostępne jest także graficzne narzędzie (trzeci interfejs) Oracle Data Miner, pozwalające na realizację zadań eksploracji oraz wizualną reprezentację wyników. Przeznaczone jest ono dla analityków biznesowych, którzy orientują się w badanych zagadnieniach i potrafią dobrać algorytm stosownie do posiadanych danych i oczekiwanych wyników. Może ono być także wykorzystywane przez programistów do wyboru kierunku rozwoju aplikacji poprzez wstępną analizę danych, tworzenie przykładowych modeli i wizualne sprawdzanie ich efektywności.

W ODM zostały zaimplementowane różne algorytmy pozwalające na tworzenie modeli eksploracji. Jak wcześniej wspomniano, możemy podzielić je na dwie grupy: z nauczycielem i bez nauczyciela. Pierwsza z nich, zawiera algorytmy realizujące funkcje:

- klasyfikacja: naiwny klasyfikator Bayesa (Naive Bayes), adaptacyjna sieć Bayes (Adaptive Bayes Network), SVM (Suport Victor Machine), indukcja drzew decyzyjnych (Decision Tree)
- regresja: SVM (Suport Victor Machine)
- ważność atrybutów: minimalna długość opisu (Minimum description Length)

W grupie drugiej, algorytmów bez nauczyciela, można wyróżnić:

- analiza skupień: algorytm k-średnich (k-Means), O-Cluster
- reguły asocjacji: apriori
- ekstrakcja cech: NMF (Non-Negative Matrix nFactorization)

Ze względu na specyfikę omawianych w artykule metod eksploracji zostaną przedstawione pokrótce algorytmy klasyfikacji, jako przykład najpopularniejszej metody uczenia nadzorowanego. W klasyfikacji wyróżniamy dwa etapy. W pierwszym etapie jest analizowany dostarczony treninowy zbiór danych. W zależności od wybranego algorytmu przeprowadzane są różne obliczenia. Wynikiem tych działań jest utworzenie modelu, który na podstawie wartości zbioru cech wejściowych pozwala na określenie wartości atrybutu przewidywanego. W drugim etapie, aby określić efektywność przewidywań, model jest testowany. Jeżeli wyniki nie są zadowalające, można powtórzyć etap tworzenia modelu z uwzględnieniem zmiany wartości jego parametrów. Do analizy dane muszą być w odpowiedni sposób przygotowane. Dane wykorzystane do testowania czy przewidywania muszą mieć identyczną postać jak dane zastosowane podczas budowania. Jednym ze sposobów przygotowania danych jest dyskretyzacja (koszykowanie). Polega na grupowaniu podobnych danych, aby wprowadzić niepowtarzalność cech danego atrybutu. W przypadku koszykowania numerycznego (dla atrybutów o wartościach liczbowych) możemy określić jeden z dwóch rodzajów:

- equal width; założmy, że min jest wartością minimalną wśród wartości danego atrybutu, max wartością maksymalną, N ilością koszy o równej długości. Przedział $[\min, \max]$ jest dzielony na N podprzedziałów o długości d , gdzie d wynosi $((\max - \min) / N)$. Kosz 1 to przedział $[\min, \min + d)$, kosz 2 to: $[\min + d, \min + 2d)$, kosz N to: $[\min + (N - 1) * d, \max]$. W takim przypadku różne kosze będą zawierały różne liczby atrybutów. Mogą się także zdarzyć kosze w których nie będzie żadnych wartości.
- quantile; założmy, że min jest wartością minimalną wśród wartości danego atrybutu, max wartością maksymalną, a N ilością koszy. Przedział $[\min, \max]$ jest dzielony na M podprzedziałów, przy czym $M \leq N$, w taki sposób aby każdy podprzedział zawierał równą liczbę elementów.

W przypadku koszykowania kategoriowego (dla atrybutów o wartościach tekstowych) występuje metoda *Top N*. Założmy, że N jest liczbą koszy. W takim przypadku zostaje wyznaczonych N najczęściej występujących wartości danego atrybutu. Atrybut jest dzielony na N+1 zbiorów, gdzie

zbiór pierwszy zawiera wartość najczęściej występującą, zbiór N zawiera wartość najmniej powtarzającą się, natomiast zbiór $N+1$ zawiera pozostałe wartości atrybutu (o mniejszej liczbie wystąpień).

W algorytmie Naive Bayes do przewidywania wartości atrybutu wykorzystuje teorię Bayesa opartą o prawdopodobieństwo warunkowe hipotezy h_i przy zaobserwowaniu danych D jako $P(h_i|D)=P(D|h_i)P(h_i)/P(D)$ [Mor05]. „Naiwność” tej metody polega na założeniu, że rozkład cech jest niezależny.

Kolejny algorytm, adaptacyjna sieć Bayesa, to szybki (ang. fast), skalowalny (ang. scalable), bezparametryczny (ang. non-parametric) sposób na przewidywanie wartości określonego atrybutu z danego zbioru danych. Bezparametrowe techniki statystyczne pozwalają na uniknięcie przypuszczeń, że populacja jest charakteryzowana przez rodzinę prostych modeli (ang. simple distributional models) takich jak standardowa regresja liniowa, gdzie poszczególni członkowie grupy są rozróżniani przez nieliczny zbiór parametrów. Jedną z przewag tego modelu nad Naive Bayes jest tworzenie reguł zrozumiałych dla umysłu ludzkiego. Osoby takie jak analitycy biznesowi, dzięki czytelnej odpowiedzi rozumieją powody zaprognozowanych wartości, daje to możliwość przekazywania zdobytej wiedzy innym [Ora05].

Algorytm SVM może być wykorzystany zarówno w klasyfikacji jak i regresji. Na podstawie danych wejściowych budowany jest liniowy model. Następnie klasy docelowe są oddzielane jak największym możliwym marginesem. W zależności od wyboru funkcji i parametrów jądra generowane są różne granice, wpływające na decyzję odpowiedniego wyboru klasy. SVM jest stosowany przy klasyfikacji tekstu, rozpoznawaniu pisma ręcznego, klasyfikacji obrazów czy analizach biometrycznych [Ora05].

Ostatnim oferowanym przez ODM algorytmem klasyfikacji jest indukcja drzew decyzyjnych. Wygenerowany model ma postać drzewa, w którym testy przeprowadzane na atrybutach zapisywane są w formie węzłów, wyniki testów w formie gałęzi, natomiast przypisanie do klasy określone w formie liści. Zaletą tego modelu jest jego przejrzystość, pozwalająca użytkownikom biznesowym zrozumieć podstawy jego działania.

ODM udostępnia różne metody pozwalające na ocenę efektywności klasyfikatora. Pierwszym narzędziem jest macierz pomyłek (ang. confusion matrix). Pozwala ona na porównanie wartości rzeczywistych z wartościami przewidzianymi. Wartości rzeczywiste pochodzą ze zbioru testowego, w którym przewidywana wartość atrybutu jest znana. Innym narzędziem do oceny modelu jest wykres przyrostu (ang. lift chart), pozwalający na graficzne podsumowanie wykorzystania modelu do wartości przewidywanej. Widać na nim, o ile częściej w stosunku do całego zbioru danych, przypadki należące do badanej klasy występują w podzbiorach danych zawierających frakcje przypadków (10%, 20% itd.) o największym, wynikającym z modelu prawdopodobieństwie przynależności do tej klasy. Efektywność klasyfikatora można także obserwować na krzywej ROC (ang. Receiver Operating Characteristics). Reprezentuje ona zależność liczby poprawnych przewidywań klasy poszukiwanej (ang. true positive) do liczby niepoprawnych przewidywań tej klasy (ang. false positive). Poprzez analizę różnych punktów na otrzymanym wykresie można dobrać odpowiedni próg prawdopodobieństwa.

4. Eksploracja danych w wykrywaniu komórek

Analizując literaturę poświęconą tematyce eksploracji danych, można zauważyć zastosowanie tego narzędzia głównie w dziedzinie CRM (ang. Customer Relationship Management). Tworzone w takim przypadku modele predykcyjne pozwalają na wdrożenie rezultatów predykcji do przeprowadzania planowanych działań handlowych czy marketingowych. Analizując dane klientów, ich preferencje i oczekiwania, można planować różne akcje. Z kolei późniejsze rejestrowanie zachowań klientów pozwala na uczenie modelu i udoskonalanie go do dalszych działań.

Wśród dostępnych publikacji niewiele miejsca poświęca się zastosowaniu eksploracji danych w innych dziedzinach, jak przemysł czy medycyna. Zachowanie takie może być związane z pewną ostrożnością. W tematyce marketingu łatwiej można zaryzykować, podjąć się eksperymentów niż w procesie produkcji czy ocenie zdrowia. Mimo tych przeciwności, można znaleźć pewne obszary zastosowań, gdzie warto podjąć chociażby próby opracowania i wdrożenia algorytmów eksploracji danych. Jednym z takich obszarów może być właśnie rozpoznawanie komórek rakowych [Per06].

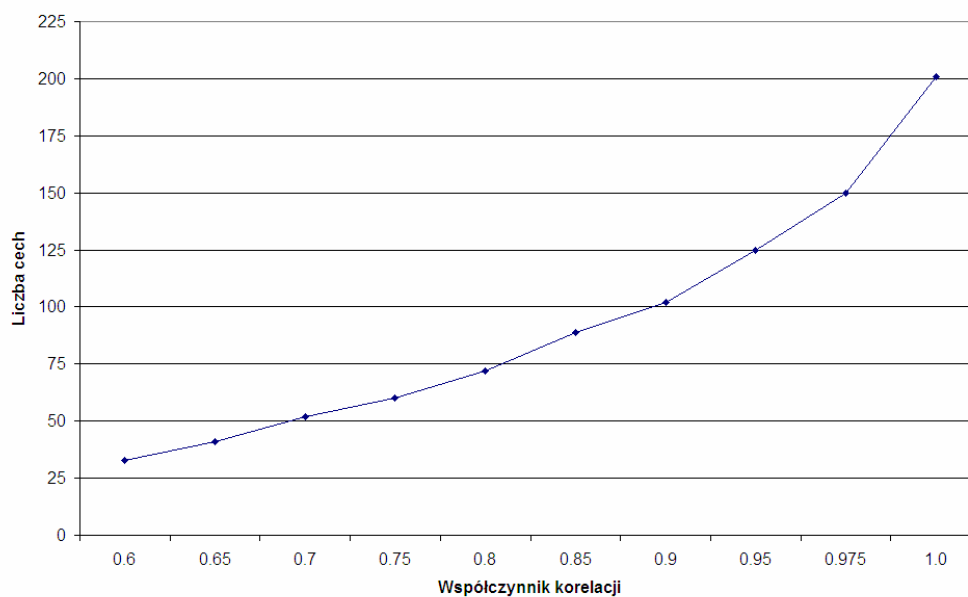
W artykule zaprezentowano wyniki wstępnych analiz, wykorzystujące algorytmy klasyfikacji Data Mining. Dostępne systemy skaningowe badające różnorodne próbki biologiczne dostarczają znacznych ilości danych. Modele wygenerowane na ich podstawie mogą okazać się pomocne do zautomatyzowania procesu wczesnego diagnozowania nowotworów. Celem przeprowadzanych badań analitycznych jest opracowanie algorytmów, które pozwolą na wytypowanie komórek rakowych na podstawie odpowiednio przygotowanych preparatów.

Rozpoznawanie komórek nowotworowych w systemie Metafer odbywa się w czterech etapach:

1. skanowanie preparatu przy różnej długości fali świetlnej,
2. analiza uzyskanego obrazu w celu wyodrębnienia poszczególnych obiektów-komórek i wyznaczenie ich cech morfometrycznych (212 cech),
3. selekcja komórek ze względu na cechy morfometryczne,
4. wyodrębnienie komórek rakowych na podstawie wprowadzonych znaczników genetycznych.

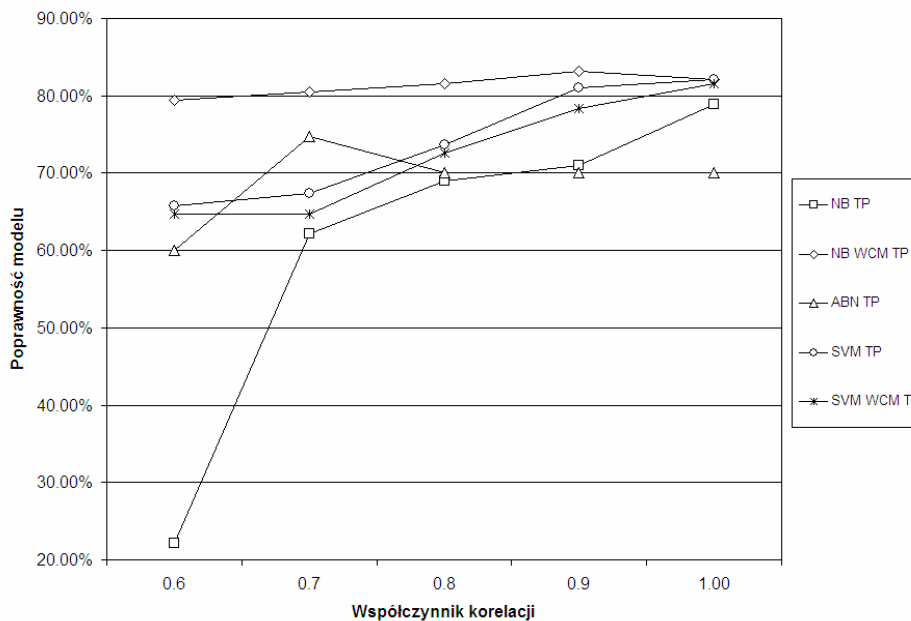
Do rozwiązania problemu wykorzystano narzędzie Data Mining Oracle 10g Release 2. Analiza zbiorów danych przeprowadzana jest dla czterech wcześniej wspomnianych algorytmów klasyfikacji. Zbiór przeznaczony do tworzenia modelu zawiera 6843 obiekty, z czego 190 zostało przez eksperta zidentyfikowanych jako komórki rakowe. Zbiór ten jest opisany poprzez 217 atrybutów. Każdy z nich odpowiada pewnej cesze komórki. Przy znacznej liczbie cech charakteryzujących poszczególne obiekty, na podstawie danych pomiarowych można zaobserwować znaczną korelację między niektórymi z nich (dla przykładu, średnica obiektu jest silnie skorelowana z polem powierzchni, gdyż poszczególne obiekty posiadają kształty zbliżone do koła). Z tego powodu kluczowym zadaniem jest problem selekcji cech w celu wyboru ich takiego podzbioru, który będzie mógł być podstawą do właściwej klasyfikacji obiektów. W artykule zastosowano metodę eliminacji cech na podstawie macierzy korelacji między poszczególnymi cechami (macierz o rozmiarach 217 x 217). Eliminowano te cechy, których współczynnik korelacji przyjmował wartość większą od założonej wartości progowej (wartości progowe współczynnika korelacji były następujące: 0.6, 0.7, 0.8, 0.9, 1.0). Na rys. 4 przedstawiono zależność liczby cech pozostawionych do klasyfikacji od granicznej wartości współczynnika korelacji.

Ze względu na możliwość wykorzystania w algorytmach macierzy kosztów, każdy z klasyfikatorów testowany był dwukrotnie. W przypadku algorytmu Adaptive Bayes Network analiza była przeprowadzana także dla każdego z trybów pracy. Do prezentacji wyników, w artykule została wybrana opcja „Multi Feature”. Dla żadnej z progowych wartości współczynnika korelacji algorytm indukcji drzew decyzyjnych nie przyniósł oczekiwanych efektów. Ilość prawidłowo odgadniętych obiektów klasy pozytywnej była zawsze zero. Dlatego, klasyfikator ten nie został uwzględniony w zestawieniu wyników w dalszej części artykułu.



Rys. 4. Zależność liczby cech od współczynnika korelacji

Wygenerowane modele poddano testowaniu. Na podstawie otrzymanych macierzy pomyłek wyznaczono procentową przewidywalność algorytmu w danej klasie (rys. 5, rys.6.).

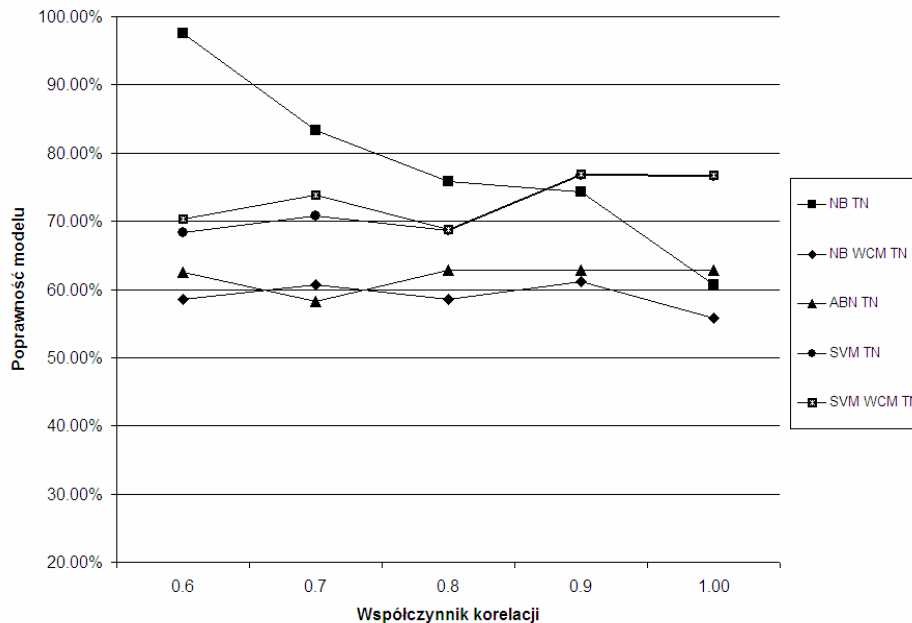


Rys. 5. Efektywność algorytmów dla klasy pozytywnej

Dla atrybutów o współczynniku korelacji mniejszym niż 0.6 najlepszy okazał się klasyfikator Naive Bayes z wykorzystaniem macierzy kosztów. Jego efektywność wyniosła prawie 80% i dla większych wartości współczynnika korelacji niewiele się zwiększyła. Drugi klasyfikator, także wykorzystujący algorytm Naive Bayes, ale już bez macierzy kosztów, na poziomie korelacji 0.6 przyniósł najgorsze efekty (22%). Wszystkie modele, z wyjątkiem opartego na sieci adaptacyjnej miały tendencję wzrostową swojej efektywności przy wzroście wartości granicznej współczynnika

korelacji. Prawie wszystkie zbiegły się na tym samym poziomie. Inną tendencję miał model adaptacyjny. Na początku zwiększył swoją efektywność, a w kolejnych analizach zmniejszył do poziomu 70% i pozostawał bez zmian.

W przypadku większości modeli można zauważyć, że dodatnia zmiana efektywności dla klasy pozytywnej skutkowałą ujemną zmianą klasy negatywnej. Najwyraźniej, zależność tą widać na modelu Naive Bayes bez macierzy kosztów. Początkowo efektywność wynosiła 97%, a ostatecznie spadła do 60%. Pewne zachwianie nastąpiło w modelach wykorzystujących SVM. Tutaj wraz ze wzrostem klasy pozytywnej rosła (przy niewielkim zachwianiu) klasa negatywna. W rezultacie była to najwyższa efektywność.

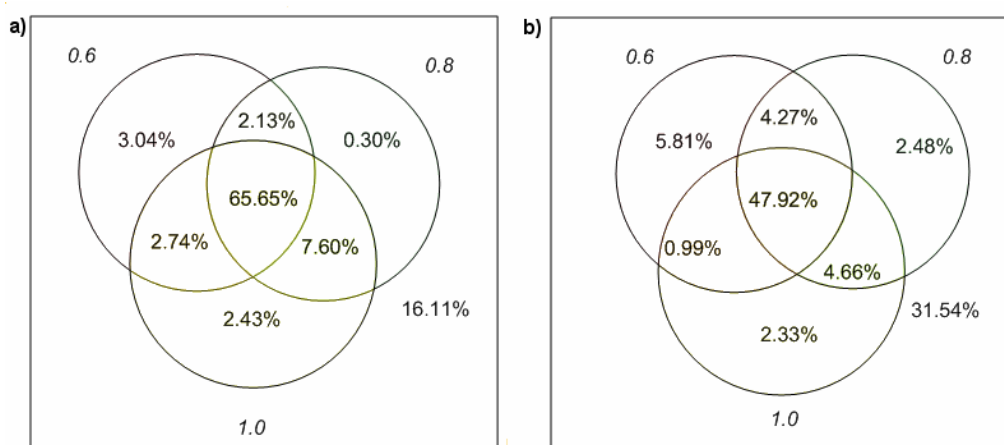


Rys. 6. Efektywność algorytmów dla klasy negatywnej

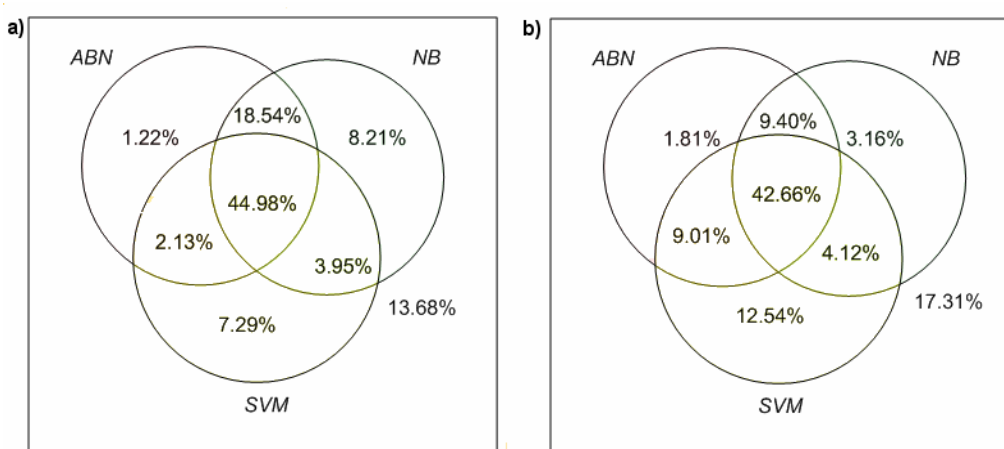
Patrząc na takie zmiany efektywności pojawia się pytanie, czy dany klasyfikator dla różnych współczynników korelacji wykrywa zawsze te same komórki. Na rys. 7 przedstawiono zestawienie wyboru obiektów dla różnych modeli tego samego algorytmu (Naive Bayes). Do zestawienia wybrano modele dla współczynników 0.6, 0.8, 1.0. Na rys. 7a widać, że wszystkie modele zgodziły się w 65.65% przypadków dotyczących stwierdzenia „dany obiekt jest komórką rakową”. Model dla współczynnika 0.6 wykrył prawidłowo 3.04% komórek rakowych, które pozostałe modele zaklasyfikowały jako zwykłe komórki. W danej klasie pozostało jeszcze 16.11% zbioru komórek rakowych, które nie zostały wykryte przez żaden z analizowanych modeli. Zakładając, że najlepsza wykrywalność jest dla współczynnika 1.0, to sytuacja idealna dla przedstawionych zbiorów byłaby gdy największa liczba znajdowała się we wspólnej części wszystkich zbiorów, a dwie w kolejne z zbiorze 1.0 i w części wspólnej 0.8 i 1.0. Pozostałe części zbiorów powinny być puste. W przypadku klasy negatywnej, polegającej na stwierdzeniu, że „dany obiekt nie jest komórką rakową” wszystkie modele zgodziły się w 47.92% danej klasy. Niestety wszystkie modele błędnie oceniły twierdząc, że obiekt jest komórką rakową aż w 31.54 % zbioru.

Na rys. 8 zostało przedstawione podobne zestawienie. Pokazuje ono jak pokrywają się przewidywania dla różnych algorytmów, ale na tym samym poziomie korelacji. Do analizy został wybrany współczynnik korelacji 0.8 oraz modele wykorzystujące macierz kosztów. Zgodność modeli dla klasy pozytywnej wyniosła prawie 45% (dla współczynnika 0.6: 28.88%; 1.0: 48.63%), dla negatywnej prawie 43% (dla współczynnika 0.6: 37.13%; 1.0: 44.07%). Dużą zgodność przewi-

dywań klasy pozytywnej można zauważyć w części wspólnej zbiorów ABN i NB, wynosi ona 18.54%. Niestety istnieje także zbiór komórek klasy pozytywnej, które przez wszystkie algorytmy zostały przewidziane negatywnie, wynosi on 13.68%. Dla tych wykresów, najlepsza efektywność wystąpi wówczas gdy wszystkie algorytmy będą przewidywać z jak najlepszym rezultatem. Wtedy w części wspólnej wszystkich zbiorów powinna znaleźć się jak największa wartość (najlepiej 100%), przy jak najmniejszej wartości poza zbiorami (najlepiej 0%).



Rys. 7. Porównanie wyboru obiektów dla klasy a) pozytywnej b) negatywnej



Rys. 8. Porównanie wyboru obiektów różnych modeli dla klas a) pozytywnej b) negatywnej

5. Podsumowanie

Celem badań jest projekt systemu automatycznej diagnostyki medycznej chorób nowotworowych w oparciu o narzędzia ORACLE. Przedstawione wyniki są wstępną analizą do procesu klasyfikacji komórek rakowych. Pozwalają one na porównanie efektywności zastosowanych algorytmów. Zabiegi takie dają możliwość zapoznania się z tak dużym zbiorem danych i nakreślają kierunki dalszych badań. Do tworzenia modelu, ze zbioru treningowego był wybierany określony podzbiór atrybutów, zależny od granicznej wartości współczynnika korelacji. W trakcie dalszych prac należy poddać analizie inne metody selekcji cech. Do podstawowych należy zaliczyć: teorię zbiorów przybliżonych (ang. rough sets) oraz macierz rozróżnialności (ang. discernibility matrix),

ranking cech, algorytmy genetyczne, itd. Przed przystąpieniem do budowania modelu, dane treningowe były przygotowywane z zastosowaniem dyskretyzacji (koszykowania). Wszystkie cechy były dzielone na 10 przedziałów (koszy). Na generowany model istotny wpływ ma także macierz kosztów. Można więc podjąć próbę modyfikacji tej macierzy. Należy także podjąć próby zastosowania innych metod klasyfikacji, jak: sieci neuronowe, zbiory rozmyte.

Bibliografia

- [AbboCD] Abbot Molecular Diagnostics: Vysis UroVysion Molecular Cytology In Detection of Bladder Cancer Recurrence, CD presentation
- [AbboHt] Abbott Laboratories Inc., <http://www.urovysion.com>
- [BCBK03] Borkowska E., Constantinou M., Binka-Kowalska A., Kałużewski B.: Diagnostyka raka pęcherza moczowego przy użyciu metody MSSCP (eksony 5-8 genu P53) i testu UroVysion, I Konferencja Użytkowników DNA Pointer System, Warszawa, 2003
- [BoSi] Borówka A., Siedlecki P.: Nowotwory układu moczowo-płciowego, Opracowanie przygotowane przez zespół ekspertów Polskiego Towarzystwa Urologicznego i Polskiego Towarzystwa Onkologii Klinicznej
- [Bro01] Brown T.A.: Genomy, Wydawnictwo Naukowe PWN, Warszawa 2001, ISBN 83-01-13439-9
- [CPS00] Cios K.J., Pedrycz W., Swiniarski R.W.: Data Mining Methods for Knowledge Discovery, Kluwer Academic Publisher Group, 2000, ISBN 0-387-33333-9
- [Guz05] Guz T.: Poprawa efektywności klasyfikatora „Box Classifier” w systemie „Metafer”, XIII Konferencja „Sieci i Systemy Informatyczne”, Łódź, 2005.
- [Mor05] Morzy M.: Oracle Data Mining – odkrywanie wiedzy w dużych wolumenach danych, XI Konferencja PLOUG Kościelisko, Październik 2005
- [Mor99] Morzy T.: Eksploracja danych: problemy i rozwiązania, V konferencja PLOUG Zakopane, Październik 1999
- [Ora05] Oracle® Data Mining; Concepts; 10g Release 2 (10.2); B14339-01; June 2005
- [Per06] Perner P.: Intelligent data analysis in medicine—Recent advances, Artificial Intelligence in Medicine (2006) 37, pp. 1-5
- [PILo01] Plesch A., Loerch T.: Metafer – a Ultra Novel High Throughput Scanning System for Rare Cell Detection and Automatic Interphase FISH Scoring, Early Prenatal Diagnosis, Fetal Cells and DNA in the Mother, Present State and Perspectives, 12th Fetal Cell Workshop, Prague, May 2001, pp.329–339
- [ZaWi03] Zajac M., Wiśniewska M.: Zastosowanie fluoroscencyjnej hybrydyzacji in situ (FISH) w identyfikacji zmian materiału genetycznego u osób z niepełnosprawnością intelektualną, Nowiny Lekarskie, 72, 2003, 9-13

