



NIESTANDARDOWE TECHNIKI WSPOMAGANIA DIAGNOSTYKI KARDIOLOGICZNEJ

*Jerzy A. Moczko, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu,
Katedra i Zakład Informatyki i Statystyki*

Jednym z najbardziej istotnych problemów w badaniach medycznych jest problem prognostyczny, który w przypadku zmiennej wyjściowej ilościowej sprowadza się do zagadnienia regresyjnego, a w przypadku zmiennej wyjściowej jakościowej do zagadnienia klasyfikacyjnego. W celu jego rozwiązania wykorzystuje się szerokie spektrum metod analitycznych, począwszy od rozmaitych technik statystycznych, poprzez techniki zgłębiania danych (data mining), kończąc na metodach sztucznej inteligencji.

Nie istnieje pojedyncze „najlepsze” narzędzie predykcyjne – możemy mówić jedynie o optymalnym doborze narzędzia w określonej sytuacji badawczej. Każde z narzędzi należących do wymienionych klas ma swoje zalety i wady, a jednym z najważniejszych aspektów prawidłowego wyboru jest analiza równoważenia obciążenia i wariancji uzyskiwanych estymatorów (*bias-variance tradeoff*). Rozważmy dla przykładu dwie techniki o całkowicie skrajnych właściwościach.

Najczęściej stosowane techniki dopasowania modelu liniowego metodą najmniejszych kwadratów bazują na bardzo silnych założeniach na temat struktury danych, co daje w wyniku stabilne, ale często mało dokładne oszacowania. Dla przykładu regresja wieloraka wymaga spełnienia aż siedmiu założeń, wymienionych między innymi w znakomitym opracowaniu Stanisza (Stanisz 2007). Pogwałcenie jakiegokolwiek z nich wpływa na jakość uzyskanego modelu predykcyjnego, aczkolwiek w różnym stopniu. W przeciwieństwie do technik dopasowania liniowego metoda k-najbliższych sąsiadów wymaga bardzo słabych założeń na temat struktury danych – predykcja jest często trafna, lecz mało stabilna. Natura wspomnianych założeń zależy od przyjętej metryki. W przypadku modelowania liniowego zakładamy m.in. globalną liniowość związku między badanymi zmiennymi, podczas gdy w przypadku stosowania metody k-najbliższych sąsiadów jedynie lokalną niezmienniczość funkcji. W efekcie modele liniowe dają gładkie kształty granic obszarów decyzyjnych, co skutkuje stosunkowo wysokim obciążeniem i niską wariancją estymatorów. Przeciwny efekt (niskie obciążenie, wysoka niestabilność estymatorów) obserwowany jest w przypadku metody k-najbliższych sąsiadów. Jakość dopasowania modelu jest określona przez tzw. oczekiwany błąd predykcji, znany również pod nazwą błędu uogólnienia lub błędu testowego. Można by sądzić, że mając wystarczająco dużą liczebność danych, na podstawie których dopasowujemy model, ta ostatnia metoda uzyska zdecydowaną przewagę nad modelami liniowymi, gdyż zwiększając wartość parametru k ,



będziemy w stanie zmniejszyć niestabilność estymatorów. Ta intuicyjna przesłanka okazuje się niestety fałszywa z uwagi na „po cichu” poczynione założenie jednorodności rozkładu danych w wielowymiarowej przestrzeni rozpiętej na użytych w badaniu predyktorach. Im większa jest wymiarowość przestrzeni, w której dokonujemy dopasowywania modelu, tym mniej mamy danych w rozpatrywanej „jednostce objętości”, przez co uśrednianie staje się coraz mniej efektywne i wariancja estymatora wzrasta. Efekt ten znany jest matematykom pod nazwą „przekleństwa wymiarowości” (*curse of dimensionality*) (Hastie, Tibshirani i Friedman 2001).

Oba wymienione zjawiska utrudniają jednoznaczny wybór modelu prognostycznego i są przyczyną podjęcia badań opisanych w dalszej części pracy.

Opis postawionego problemu medycznego oraz charakterystyka materiału badawczego i użytych metod analitycznych

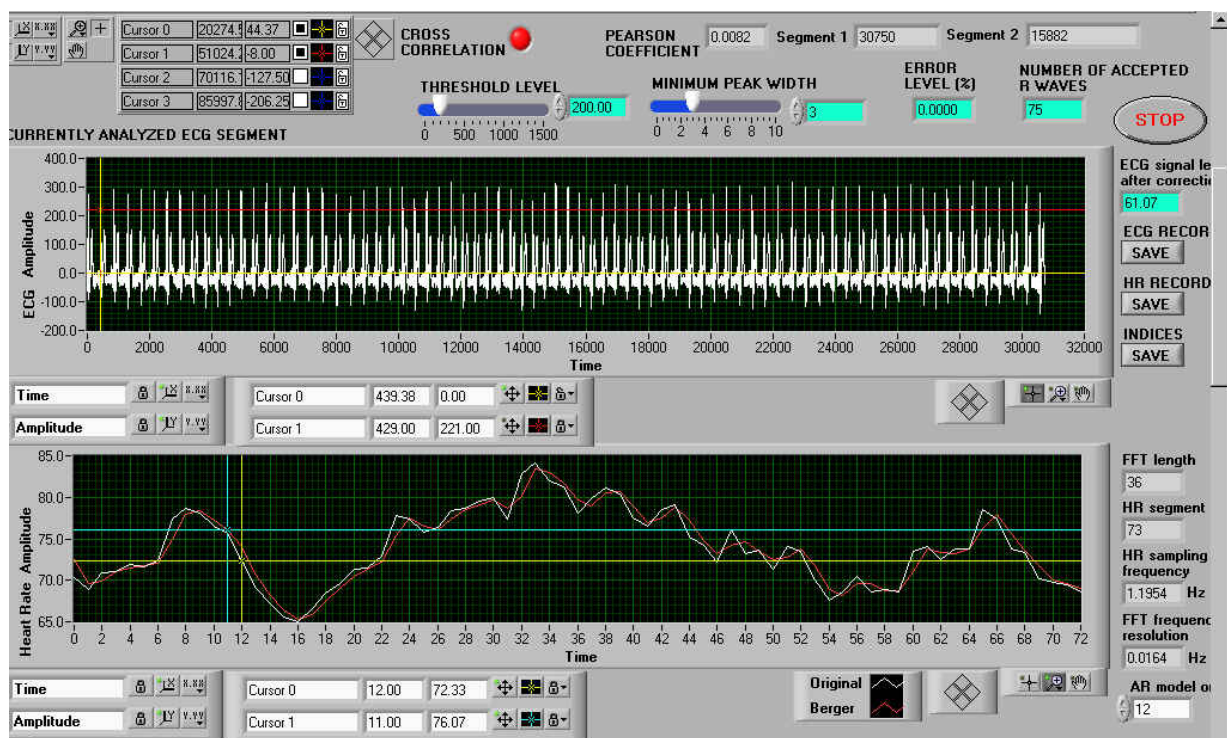
Szczególną cechą danych biomedycznych jest ich wysoka zmienność zarówno między-, jak i wewnątrzsobnicza. Właściwość ta jest zasadniczym powodem tego, że powszechnie stosowane standardowe, jednowymiarowe metody statystyczne rzadko dają wyniki o wysokiej jakości predykcyjnej (Moczko 2004). O ile do zebranych danych udaje się na ogół dopasować istotne statystycznie modele wiążące istniejące zależności, to z oczywistych względów jakość predykcji jest z reguły wyższa dla podzbioru danych uczących, gorsza zaś dla danych testujących. Sytuacja ta oznacza niską zdolność modeli do uogólniania wniosków.

W realizowanym badaniu podjęto próbę oceny wartości prognostycznej najczęściej wykorzystywanych przez kardiologów wskaźników zmienności krótko- i długoterminowej częstości uderzeń serca. Wszystkie mierzone są w skali interwałowej, a każdy z nich ma odmienne właściwości reagowania na różne aspekty rozkładu danych. Dla przykładu stosowany powszechnie w położnictwie i kardiologii perinatalnej i neonatalnej wskaźnik zmienności długoterminowej wg Yeha jest z punktu widzenia statystyka niczym innym, jak współczynnikiem zmienności następujących po sobie interwałów między kolejnymi załamkami R zmierzonych w okresie o długości 30 sekund; analogiczny wskaźnik zmienności długoterminowej wprowadzony przez de Haana stanowi rozstęp międzykwartyłowy modułów odległości między kolejnymi punktami leżącymi na dwuwymiarowej płaszczyźnie ($T_{RR}(i)$, $T_{RR}(i - 1)$) utworzonej przez czasy trwania kolejnych 128 interwałów R-R. Przez lata prowadzonych badań uczeni wprowadzili dziesiątki wskaźników opisujących różne aspekty pracy serca w rozmaitych dziedzinach (czasu, częstotliwości, połączonych dziedzinach czasu i częstotliwości) i badali związki każdego z nich z osobna ze stanem klinicznym pacjentów. Jak wspomniano poprzednio, rozmaite właściwości matematyczne poszczególnych wskaźników powodują niejednakową ich reakcję na zmiany zachodzące w sekwencjach interwałów R-R. Powstaje zatem pytanie, czy tworząc odpowiednią kombinację znanych wskaźników, można uzyskać lepszą zdolność predykcyjną od tej uzyskiwanej na podstawie każdego ze wskaźników użytego z osobna. Prace zrealizowano, wykorzystując zaprojektowaną i wdrożoną do użytku bazę danych koronarograficznych. Zawiera ona informacje o 358 parametrach wyrażonych w rozmaitych skalach pomiarowych



uzyskanych od 550 pacjentów. Szczegółowa struktura bazy opisana jest w pracy (Moczko 2002). Równoległe z danymi alfanumerycznymi rejestrowane były zapisy elektrokardiograficzne przy użyciu aparatury MEDEA w 12 odprowadzeniowym systemie Francka. U każdego pacjenta zapisy zostały wykonane co najmniej dwukrotnie (przed i po koronarografii). Analizy zapisu EKG oraz częstości uderzeń serca w dziedzinach czasu, częstotliwości i połączonych dziedzinach czasu-częstotliwości dokonano przy użyciu zaprojektowanych w Katedrze Informatyki i Statystyki Uniwersytetu Medycznego narzędzi wirtualnych (LabVIEW,® National Instruments, Inc.) (Moczko 2002). Fragmenty płyt czołowych tych narzędzi pokazano na rysunkach 1 i 2. Z tak uzyskanych danych wytworzono w programie *STATISTICA Data Miner* zbiory uczący i testujący i wykorzystano je do realizacji zadań klasyfikacyjnych mających na celu udzielanie odpowiedzi na definiowane przez kardiologów zapytania. Zadania te obejmują 20 modeli wielowymiarowych wykorzystujących następujące procedury matematyczne:

1. analiza dyskryminacyjna,
2. analiza logistyczna,
3. analiza bayesowska,
4. drzewa klasyfikacyjno-regresyjne (6 modeli),
5. technika wielozmiennej regresji adaptacyjnej z użyciem funkcji sklejaných (MARSpline) (3 modele),
6. technika wektorów nośnych (Support Vector Machines – SVM) (2 modele),
7. technika k-najbliższych sąsiadów (6 modeli).



Rys. 1. Fragment panelu narzędzia wirtualnego wybierającego żądany fragment zapisu EKG.



Rys. 2. Fragment panelu narzędzia wirtualnego raportującego wyniki analizy wybranego fragmentu zapisu EKG w dziedzinach czasu, częstotliwości i połączonych dziedzinach czasu-częstotliwości.

W przeciwieństwie do powszechnie stosowanych metod prostych, wielowymiarowe techniki badawcze (statystyczne, data-mining) pozwalają na wykrycie i wyeliminowanie zmiennych redundantnych, wikłających, mediacyjnych oraz efektów interakcyjnych. Ma to zasadniczy wpływ na zbudowanie prawidłowego modelu prognostycznego. Niestety interpretacja modeli wielowymiarowych jest z natury rzeczy o wiele bardziej złożona i czasami bywa wręcz sprzeczna z ogólnie znanymi, potwierdzonymi wielokrotnie w badaniach jednowymiarowych faktami. Przykład takiej sytuacji przytacza w swej pracy Hastie i Tibshirani (Hastie i Tibshirani 1987), omawiając przykład dopasowania logistycznego modelu regresyjnego do badań kardiologicznych CORIS prowadzonych na początku lat osiemdziesiątych w Południowej Afryce. Brak w wytworzonym modelu istotności statystycznej wpływu ciśnienia skurczowego krwi i otyłości na ryzyko zapadalności na niedokrwienne chorobę serca dla niejednego kardiologa będzie herezją. Dopiero zrozumienie faktu wewnętrznego skorelowania między włączonymi do modelu predyktorami (wiek, poziom LDL, uwarunkowania genetyczne, spożycie alkoholu, palenie tytoniu, ciśnienie skurczowe, stopień otyłości) pozwala zrozumieć, dlaczego wytworzony model nie włącza powszechnie znanych i dokładnie zbadanych zmiennych opisujących. Są one niepotrzebne (nawet niekiedy szkodliwe) z chwilą obecności w modelu innych, silnie z nimi skorelowanych czynników. Dlatego też koncepcja budowy modelu na bazie włączania do niego takich zmiennych opisujących, które są dobrymi predyktorami w modelach prostych często kończy się niepowodzeniem.

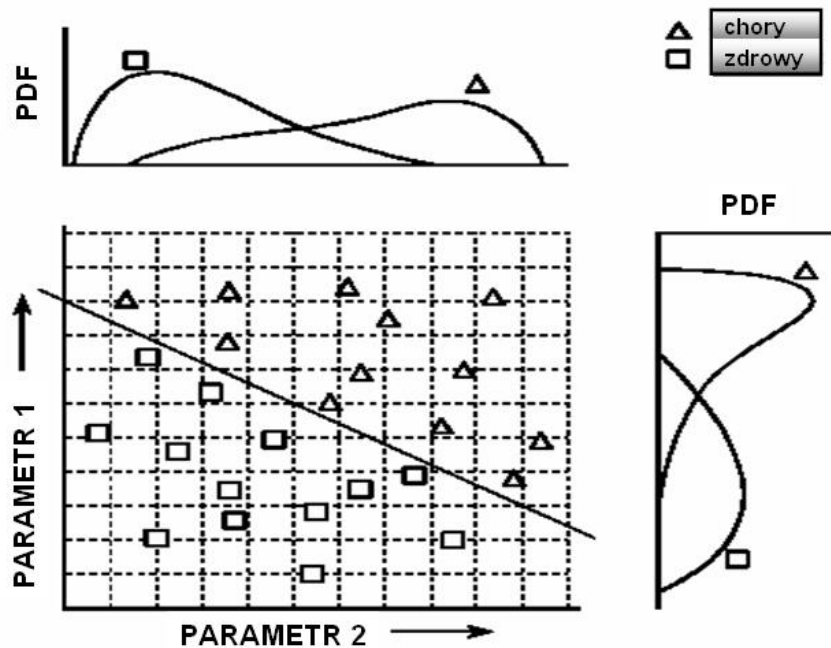


Przykład realizacji metaklasyfikatora w środowisku pakietu statystycznego STATISTICA

W badaniach medycznych wykorzystuje się informacje o charakterze alfanumerycznym, sygnałowym lub obrazowym. Ekstrahuje się z nich deskryptory charakterystyczne dla poszczególnych stanów badanych obiektów i wyraża w jednej ze skal pomiarowych: ilorazowej, interwałowej, porządkowej lub nominalnej. Ta różnorodność skal pomiarowych powoduje, że przy przetwarzaniu zebranych informacji stosuje się szerokie spektrum technik analitycznych bazujących na rozmaitych założeniach co do rozkładu analizowanych danych.

W prezentowanej pracy skupiamy się na pojedynczym prognostycznym zagadnieniu klasyfikacyjnym, które spośród poddanych analizie wskaźników opisujących pracę serca można uznać za czynniki prognostyczne wystąpienia zawału serca? Jako predyktorów użyto dwadzieścia deskryptorów powszechnie używanych do oceny sygnału częstości uderzeń serca (z punktu widzenia teorii sygnałów jest to szereg zdarzeń (*time events*) a nie klasyczny szereg czasowy (*time series*)): SDNN, RMSSD, wskaźniki zmienności długo- i krótkoterminowej wg de Haana, wskaźniki zmienności długo- i krótkoterminowej wg Yeha, wskaźniki zmienności długo- i krótkoterminowej wg Hueya, wskaźniki zmienności długo- i krótkoterminowej wg Zugaiba, wskaźniki zmienności długo- i krótkoterminowej wg Daltona, wskaźnik Allana, szerokość wstęgi oscylacyjnej, wskaźniki spektralne VLF, LF, HF oraz LF/HF. Szczegółowe definicje matematyczne wymienionych deskryptorów znajdzie Czytelnik w pracy (Moczko 2002). Zmienna opisywana jest wyrażona w skali nominalnej i opisuje dwa stany: wystąpienie lub brak wystąpienia zawału.

Jak wspomniano we wstępie, proste, jednowymiarowe analizy każdego ze wskaźników z osobna najczęściej nie dają optymalnych rezultatów dyskryminacyjnych z uwagi na występujące zmienności między- i wewnątrzosobnicze. Przyjmowane przez lekarzy normy dla danego parametru reprezentowane są nie poprzez pojedyncze wartości liczbowe, lecz poprzez zakresy przedziałowe. Niestety i one mogą na siebie częściowo nachodzić, co w rezultacie prowadzi do niejednoznacznych wniosków. Na rys. 3 widzimy przykład takiej sytuacji, w której rozkłady parametrów 1 i 2 dla osób zdrowych i chorych częściowo się pokrywają, co uniemożliwia jednoznaczne podjęcie decyzji, zarówno na podstawie osobno analizowanej wartości parametru 1, jak i parametru 2. Podejście dwuwymiarowe (wnioskowanie na podstawie jednocześnie pomierzonych parametrów 1 i 2) pozwala na skonstruowanie linii jednoznacznie separującej obszar osób zdrowych i chorych. Uogólnienie tego spostrzeżenia na większą liczbę wymiarów daje nam możliwość utworzenia w k-wymiarowej hiperprzestrzeni rozpiętej na k predyktorach hiperpowierzchni rozgraniczającej interesujące nas obszary.



Rys. 3. Klasyfikacja przypadków do grupy zdrowych lub chorych na podstawie wartości dwóch mierzonych parametrów.

Niestety to proste złożenie informacji poprzez dodanie kolejnych deskryptorów może prowadzić do zaskakujących wyników z uwagi na występowanie efektów redundancji, uwikłania, interakcji, mediacji itp. (Vittinghoff, Glidden, Shiboski i McCulloch 2001). Dla przebadania opisanego problemu, wykorzystując dane z opisanej uprzednio bazy badań koronarograficznych, skupiono się na rozwiązaniu następującego zagadnienia: które kombinacje spośród poddanych analizie wskaźników opisujących pracę serca można uznać za czynniki prognostyczne wystąpienia zawału serca? W tym celu wykonano eksperymenty obliczeniowe mające na celu udzielenie odpowiedzi na następujące pytania cząstkowe:

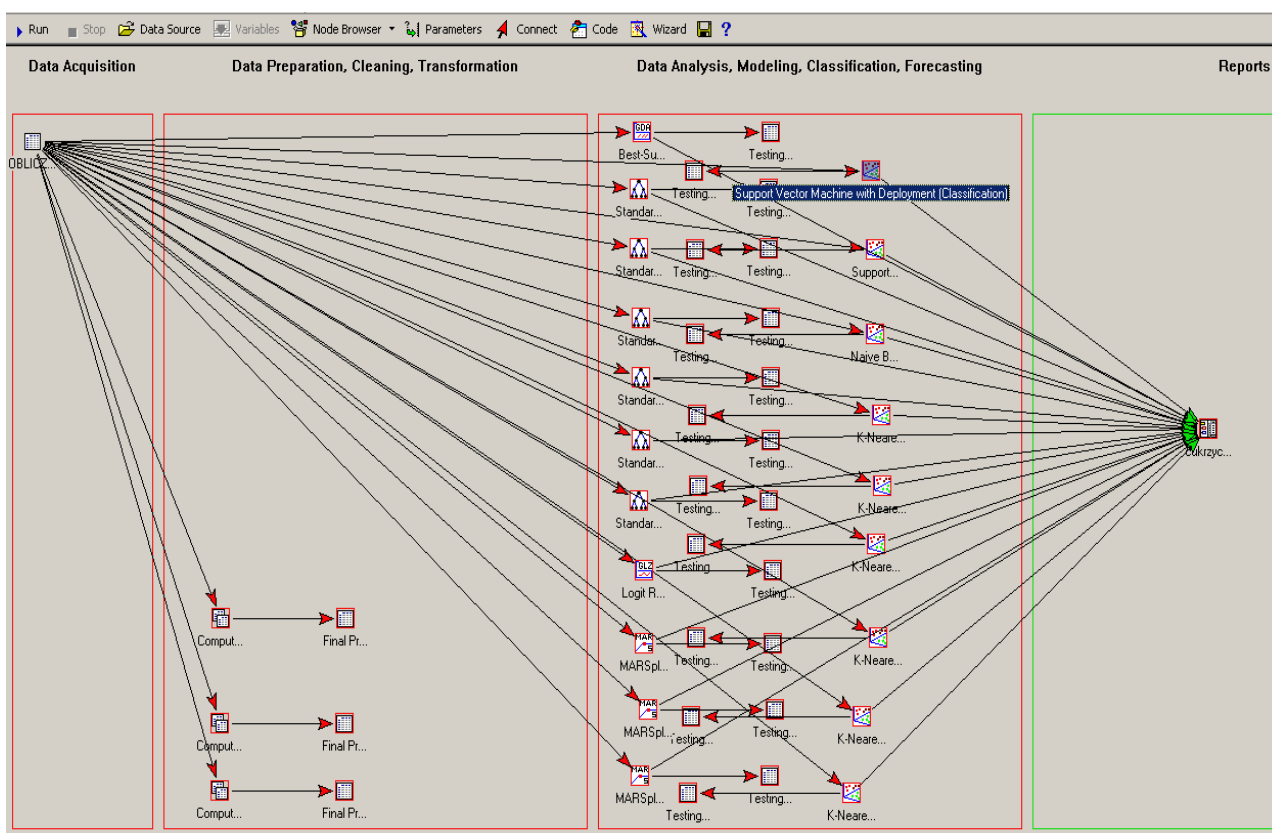
1. Które z zastosowanych technik klasyfikacyjnych dają optymalne rozwiązania klasyfikacyjne?
2. Jaki zestaw predyktorów można uznać za efektywny i reprezentatywny w konstrukcji klasyfikatora (metaklasifikatora)?
3. Jaki jest wpływ stosunku liczebności grupy testującej do grupy uczącej na jakość uzyskanej klasyfikacji?

W celu zbadania przydatności wielowymiarowych metod analitycznych do budowy klasyfikatorów lub metaklasifikatorów użyto następujących technik: analizę dyskryminacyjną, regresję logistyczną, techniki MARSplines (metody rekurencyjnego podziału przestrzeni cech do budowy modelu regresyjno-klasyfikacyjnego w postaci krzywych sklejanych), standardowe drzewa klasyfikacyjno-regresyjne CRT, metodę wektorów nośnych, naiwne klasyfikatory bayesowskie oraz metodę K najbliższych sąsiadów. W analizie dyskryminacyjnej przyjęto wartości prawdopodobieństw klasyfikacyjnych apriorycznych proporcjonalnych do wielkości podgrup. Podobne założenie zastosowano w przypadku drzew CRT.



W tych ostatnich przyjęto równe koszty błędnych klasyfikacji oraz miarę dobroci dopasowania wg Giniego. Wykorzystano kolejno trzy możliwości określenia reguły zatrzymania: przytnij przy błędzie złej klasyfikacji, przytnij przy określonej wartości błędu standardowego (dewiacji, odchylenia) oraz bezpośrednio zatrzymywanie typu FACT (z minimalną liczbą błędnych klasyfikacji równą 5, co stanowiło 1.6 % liczebności całej bazy). Każda z trzech opcji była realizowana dwukrotnie: z 10-krotnym sprawdzianem krzyżowym i bez sprawdzianu krzyżowego. W nieparametrycznej metodzie regresji MARSplines dopuszczono interakcję dwóch, czterech lub siedmiu predyktorów. W przypadku techniki wektorów nośnych (SVM – *Support Vector Machines*) budowano przestrzeń decyzyjną separującą obiekty należące do jednej z dwóch analizowanych klas przy wykorzystaniu radialnych funkcji bazowych (RBF) (parametr jądra $\gamma=0.2$) i klasyfikacyjnej funkcji błędu typu 1 z wartością pojemności równą 10. W metodzie k-najbliższych sąsiadów użyto odległości euklidesowych, a jako parametr K przyjęto kolejno wartości 1, 3 oraz 5.

W ostatniej fazie obliczeń przeprowadzono prognozę klasyfikacyjną, wykorzystując metodę kontaminacji wszystkich uzyskanych modeli. W celu porównawczym jako prognozę wybierano wynik „głosowania” spośród: wszystkich wytworzonych modeli, trzech najlepszych wytworzonych modeli oraz najlepszy uzyskany wynik. Do przeprowadzenia opisaney analizy wykorzystano moduł *Data Miner* pakietu statystycznego *STATISTICA* firmy StatSoft (<http://www.statsoft.pl/textbook/stathome.html>). Na rys. 4 przedstawiono strukturę obszaru roboczego *Data Minera* realizującą opisaną sekwencję zadań.



Rys. 4. Struktura obszaru roboczego *Data Minera* realizująca sekwencję zadań opisanych w bieżącym rozdziale.



Wyniki

Obliczenia przeprowadzono dla pięciu układów proporcji wielkości zbioru testującego do całkowitej wielkości zbioru: 10%, 20%, 30%, 40% i 50%. Z uwagi na dużą objętość uzyskanych wyników w tabeli 1 przedstawiono jedynie te wyniki, które w opcji najlepszego uzyskanego rezultatu dały najniższy odsetek błędnych klasyfikacji. W każdej sytuacji najlepszą klasyfikację zbioru uczącego, uzyskano stosując metodę wektorów nośnych bez walidacji krzyżowej, metodę wektorów nośnych z 10-krotną walidacją krzyżową oraz metodę najbliższych sąsiadów z wartością $K=1$ bez walidacji krzyżowej. Podobnie dla klasyfikacji zbioru testującego najlepsze wyniki uzyskano, stosując metodę MARSpline bez interakcji. Oznacza to, że wspomniane metody nieparametryczne dają lepszą predykcję zarówno w przypadku odtwarzania danych uczących, jak i procesu uogólniania predykcji z danymi nowymi. Mimo przeprowadzania analizy w przestrzeni 20-wymiarowej zjawisko „przekleństwa wymiarowości” nie odegrało znaczącej roli, co może świadczyć o stosunkowo jednorodnym pokryciu przypadkami badanej hiperprzestrzeni.

Tabela 1. Wyniki pomiarów (0 – zawał nie wystąpił, 1 – wystąpienie zawału).

Proporcja wielkości zbioru testującego	Wartości rzeczywiste	Zbiór uczący Wartości predykcji		Zbiór testujący Wartości predykcji	
		0	1	0	1
10%	0	175	0	21	1
	1	0	66	4	3
20%	0	164	0	31	3
	1	0	54	14	5
30%	0	137	0	59	2
	1	0	53	15	5
40%	0	112	0	81	5
	1	0	48	14	11
50%	0	100	0	94	4
	1	0	37	28	8

Dyskusja

Zastosowanie do analizy tych samych danych siedmiu rozmaitych technik analitycznych w łącznie 20 konfiguracjach pozwala na wszechstronną ocenę sformułowanego problemu klasyfikacyjnego. Użyte metody i ich poszczególne warianty, posługując się różnymi założeniami, badają rozmaite właściwości przedstawianych danych. Dla przykładu użyty algorytm analizy dyskryminacyjnej może być stosowany tylko i wyłącznie dla zbioru predyktorów pomierzonych w skali interwałowej lub ilorazowej. Zakłada on, że dane mają rozkład wielowymiarowy normalny, zachodzi jednorodność macierzy wariancji/kowar-



iancji, brak jest skorelowania średnich i wariancji oraz macierz, której kolumny tworzą poszczególne predyktory, jest dobrze uwarunkowana. Nieduże odstępstwa od pierwszych dwóch założeń nie powodują istotnych zmian we wnioskowaniu, natomiast pogwałcenie trzeciego założenia stanowi silne zagrożenie dla trafności predykcji. Sytuacja ta występuje najczęściej wtedy, gdy w niektórych grupach wyników występują punkty odstające. W przypadku gdy pośród predyktorów pojawiają się wielkości pomierzone w skali porządkowej lub nominalnej, prosta analiza dyskryminacyjna musi zostać zastąpiona uogólnionymi modelami analizy dyskryminacyjnej. Zaletą tej metody jest, że jako metoda parametryczna bazująca na założeniu globalnej liniowości istniejących związków stosunkowo łatwo obchodzi problemy związane z efektem zjawiska „przekleństwa wymiarowości”.

W przypadku prostej analizy logistycznej przewidujemy na podstawie predyktorów ilościowych lub jakościowych zmienną ciągłą, która zawiera się w granicach 0–1. Uogólniony model regresji logistycznej, będący naturalnym rozwinięciem prostego modelu logistycznego dla odpowiedzi binarnych, pozwala na uzyskiwanie zmiennej zależnej wartości z pewnego przedziału. Ogólnie metody parametryczne mają przewagę w sytuacjach danych rzadkich, z dużym udziałem szumu pomiarowego i ze stosunkowo małą liczbą przypadków uczących.

O ile w modelu analizy dyskryminacyjnej i logistycznej zakładaliśmy konkretny kształt funkcji wiążącej (odpowiednio liniowy i sigmoidalny), o tyle technika MARSplines reprezentuje przykład nieparametrycznych metod regresyjno-klasyfikacyjnych, w których nie zakłada się konkretnej zależności funkcyjnej wiążącej predyktory ze zmienną zależną, lecz model zależności budowany jest przy wykorzystaniu iloczynów funkcji bazowych o określonym kształcie oraz współczynników oszacowanych na podstawie wartości predyktorów i ich interakcji. Innymi słowy, ogólny mechanizm działania MARSplines należy traktować jako wielokrotną, odcinkową regresję liniową. Aczkolwiek samo sformułowanie tej techniki łączy się ściśle z problemem regresyjnym, to można ją rozszerzyć na problemy klasyfikacyjne poprzez przyporządkowanie badanego obiektu do klasy o największej wartości prognozowanej odpowiedzi. Jak widać z przytoczonych wyników, metoda ta okazała się najbardziej elastyczna i efektywna dla rozwiązania postawionego binominalnego problemu klasyfikacyjnego.

Podobnie do metod nieparametrycznych możemy zaliczyć metodę k-najbliższych sąsiadów, w której stosujemy pamięciowy model określony przez zespół „przykładowych” przypadków uczących o znanych wartościach zmiennej zależnej. W metodzie tej zespoły predyktorów, jak i zmiennej opisywanej mogą być określone w dowolnej skali.

Nieliniowe granice decyzyjne konstruuje również (jednakże na całkowicie innej zasadzie) niezwykle elastyczna metoda wektorów nośnych. Wektor nośny będący zbiorem punktów w przestrzeni cech (uzyskanych z przetransformowania danych wejściowych przy użyciu różnych funkcji bazowych) definiuje hiperprzestrzeń odpowiadającą poszczególnym klasom. W rozważanym zadaniu metoda ta okazała się najlepszą w odtwarzaniu danych uczących.

Naiwny klasyfikator Bayesa oparty na znanym wzorze Bayesa jest szczególnie prostym i często efektywnym narzędziem klasyfikacyjnym, lecz niestety opiera się on na najczęściej słabo spełnionym warunku niezależności statystycznej predyktorów. Mimo to jest on



uznanym narzędziem w przypadku dużej liczby wymiarów przestrzeni predyktorów. Bardzo słabe (najgorsze ze wszystkich wyników) rezultaty klasyfikacyjne wynikają w naszej sytuacji z całkowitego pogwałcenia wspomnianego założenia.

Drzewa regresyjno-klasyfikacyjne (CRT) należą również do klasy metod nieparametrycznych opartych na znalezieniu sekwencji logicznych warunków podziału o strukturze logicznej „jeżeli-to”, prowadzących do jednoznacznego zaklasyfikowania obiektów. Wielką ich zaletą jest duża odporność na pojawiające się dane o wartościach odstających. Można je wykorzystywać do analiz z predyktorami pomierzonymi w dowolnej skali pomiarowej. Aczkolwiek z punktu widzenia łatwości interpretacyjnej są one chętnie wykorzystywane przez lekarzy, mają jednak poważną wadę wynikającą z ich dużej zmienności. Nawet niewielkie zmiany w układzie danych mogą prowadzić do całkowitej zmiany struktury drzewa. Właściwość ta wynika z hierarchicznego sposobu realizacji procesu klasyfikacyjnego, w którym wystąpienie zmiany w podziale na wyższym poziomie automatycznie narusza podziały niższego rzędu.

Metody parametryczne, które wymagają silnych założeń co do struktury danych, dają z reguły wyniki predykcyjne stabilne, lecz często nietrafne (obciążone). W przeciwieństwie do nich, metody nieparametryczne pozwalają na uzyskanie prawidłowych predykcji kosztem wysokiej niestabilności wyników [1]. Każda z zastosowanych metod prowadzi w zasadzie do innego zestawu predyktorów wchodzących w skład zbudowanego modelu, co osobom mało doświadczonym w zagadnieniach statystycznych może wydawać się dziwne, jeżeli nie całkiem podejrzane. Należy pamiętać, że sytuacja taka pojawia się dość często, nawet przy zastosowaniu pojedynczej metody badawczej. Dla przykładu, stosując wielowymiarową regresję krokową w modelu wstępującym i zstępującym, uzyskujemy zazwyczaj różne zestawy istotnych predyktorów. Jeszcze silniej zjawisko to występuje przy zastosowaniu różnych technik analitycznych, z których każda zwraca uwagę na inne właściwości badanych danych (np. miarę tendencji centralnej, rozproszenie danych, kształt rozkładu itp.). Opisany efekt jest zazwyczaj dla badacza niezwykle korzystny. Sytuacja, w której możemy decydować czy niemal identyczną jakość klasyfikacyjną uzyskamy, stosując badania mniej lub bardziej inwazyjne, dających mniejszą lub większą szansę uniknięcia ewentualnych powikłań, uzyskać wynik taniej lub drożej, nie ma nic wspólnego z niejednoznacznością czy też sprzecznością otrzymywanych rezultatów. W analizie redukcji wymiarowości o wiele bardziej niebezpieczne jest występowanie zjawisk interakcji i uwikłania w zespole zmiennych predykcyjnych. Brak merytorycznej wiedzy na temat istniejących związków przyczynowo-skutkowych może doprowadzić do wyciągnięcia całkowicie fałszywych wniosków, co znakomicie ilustrują prace (Hastie, Tibshirani i Friedman 2001) oraz (Vittinghoff, Glidden, Shiboski i McCulloch 2001).

Literatura

1. Hastie T., Tibshirani R., Nonparametric logistic and proportional odds regression, Applied Statistics vol.36: 260-276, 1987.



2. Hastie T., Tibshirani R., Friedman R., The Element of Statistical Learning. Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
3. Internetowy Podręcznik Statystyki firmy StatSoft <http://www.statsoft.pl/textbook/stat-home.html>.
4. Moczko J.A., Advanced methods of heart rate signals processing and their usefulness in diagnosis support I. Mathematical heart rate descriptors and virtual instrumentation, Computational Methods in Science and Technology, vol. 8 nr 2 s. 65-76, 2002.
5. Moczko J.A., Advanced methods of heart rate signals processing and their usefulness in diagnosis support II. Univariate statistical techniques, Computational Methods in Science and Technology, vol. 10 nr 1 s. 73-82, 2004.
6. Stanisław A., Przystępny kurs statystyki z zastosowaniem *STATISTICA PL* na przykładach z medycyny – Tom 2. Modele liniowe i nieliniowe. StatSoft Polska Sp. z o.o., Kraków 2007.
7. Vittinghoff E., Glidden D.V., Shiboski S.C., McCulloch C.E., Regression Methods in Biostatistics. Linear, Logistic, Survival and Repeated Measures Models, Springer-Verlag, New York, 2001.