



WYKORZYSTANIE DRZEW KLASYFIKACYJNYCH DO WSPOMAGANIA PROCESÓW PODEJMOWANIA DECYZJI

mgr Małgorzata Misztal⁶

Uwagi wstępne

Działalność człowieka to nieustanny proces podejmowania decyzji. Z każdą decyzją wiąże się jednak możliwość popełnienia błędu, a dodatkowo, wybór danej decyzji ze zbioru decyzji dopuszczalnych nie przesądza w sposób jednoznaczny o wyniku lub skutku podjęcia decyzji. Każda decyzja wymaga systematycznego i rozważnego zgłębienia wszystkich możliwych sposobów działania, a następnie wyboru jednego z nich. Zatem celem analizy decyzyjnej jest tworzenie procedur pozwalających na porównanie alternatywnych sposobów działania i kryteriów, na podstawie których można by podejmować decyzje.

Znaczny wpływ na powstanie i rozwój narzędzi usprawniających działalność człowieka wywarł rozwój technik komputerowych. Postępy informatyzacji stały się siłą napędową rozwoju tych kierunków nauki, które bazują na mocy obliczeniowej komputerów lub wykorzystują specyfikę obliczeń komputerowych w celu bardziej efektywnego rozwiązywania problemów występujących w praktyce. Przykładem są tutaj metody wspomagania procesów podejmowania decyzji będące przedmiotem badań w rozpoznawaniu obrazów, przy czym obraz jest tu rozumiany jako ilościowy opis obiektu, zdarzenia lub zjawiska. Rozpoznawanie jest procesem wieloetapowym, który można zdefiniować jako proces przetwarzania informacji, podczas którego relatywnie duża ilość danych wejściowych zostaje przetworzona na mniejszą ilość danych użytecznych, zakończony klasyfikacją.

Klasyfikacja to czynność przydzielenia obiektu do klas (przypisanie obiektowi numeru klasy). Cele klasyfikacji określa Gatnar (1998) następująco:

- ◆ Uzyskanie jednorodnych przedmiotów badań, w których łatwiej wyróżnić czynniki systematyczne. Rezultatem jest tutaj redukcja dużej liczby obiektów (cech) do kilku podstawowych kategorii i zmniejszenie nakładu pracy oraz czasu analiz dzięki ograniczeniu liczby danych.
- ◆ Chęć odkrycia nieznanymi struktur analizowanych danych lub porównanie obiektów wielowymiarowych.
- ◆ Możliwość weryfikacji hipotez dotyczących charakteru danych i wnioskowania o nieznanymi cechach obiektów na podstawie znajomości klas, do których należą.

Szczególnie użyteczną metodą klasyfikacji zdaje się być analiza dyskryminacji, dzięki której zbiór obserwacji można przydzielić do k klas (populacji) posiadających własność jednorodności, przy założeniu, że charakterystyki tych klas są przynajmniej częściowo znane. Charakterystyki populacji

⁶ Katedra Metod Statystycznych, Uniwersytet Łódzki.



określa się zwykle na podstawie tzw. zbioru uczącego, czyli zbioru obserwacji, co do których posiadamy informacje, do jakich klas należą.

Wadą metod dyskryminacji jest to, iż wymagają one spełnienia dwóch założeń: zmienne reprezentujące cechy obiektów muszą mieć łącznie wielowymiarowy rozkład normalny oraz macierze wariancji kowariancji w poszczególnych klasach mają być równe (przy liniowych funkcjach dyskryminacyjnych). W praktyce oba te wymogi są dużym ograniczeniem klasycznych metod dyskryminacji. Rozwiązaniem tego problemu może być zastosowanie analizy dyskryminacji bez weryfikacji założeń bądź też wykorzystanie innych metod, które nie wymagają spełnienia tak surowych warunków stosowalności. Do takich metod należą między innymi drzewa klasyfikacyjne.

Ogólna charakterystyka drzew klasyfikacyjnych

Drzewa klasyfikacyjne wykorzystywane są do określania przynależności obiektów do klas jakościowej zmiennej zależnej na podstawie pomiarów jednej lub więcej zmiennych predykcyjnych. Innymi słowy, drzewo klasyfikacyjne przedstawia proces podziału zbioru obiektów na jednorodne klasy. Podział jest dokonywany w oparciu o wartości cech obiektów, liście odpowiadają klasom, do których należą obiekty, a krawędzie drzewa reprezentują wartości cech, na podstawie których dokonano podziału (por. Gatnar 1998).

Najprostszy podział drzew klasyfikacyjnych wyróżnia drzewa binarne i niebinarne.

W drzewach binarnych z każdego węzła wychodzą jedynie dwie krawędzie – każdy zbiór obiektów dzieli się na dwa rozłączne podzbiory. Drzewa niebinarne, z kolei, mają przynajmniej jeden węzeł, z którego wychodzą więcej niż dwie krawędzie.

W przypadku drzew binarnych i obiektów charakteryzowanych przez cechy ilościowe wyróżnić można drzewa jedno i wielowymiarowe. Rozróżnienie to wiąże się z postacią testu w węzle, który pozwala na dokonanie podziału zbioru obiektów. Dla drzew jednowymiarowych testy te mają postać: $x_i < C$, a w przypadku drzew wielowymiarowych mamy do czynienia z kombinacjami liniowymi: $a_0 + \sum_i a_i x_i > 0$.

Proces tworzenia drzewa klasyfikacyjnego polega na rekurencyjnym podziale zbioru uczącego (zawierającego obiekty, co do których wiemy do jakich klas należą) na podzbiory aż do uzyskania ich jednorodności ze względu na przynależność obiektów do klas. Celem jest tu utworzenie drzewa o jak najmniejszej liczbie węzłów, aby otrzymać jak najprostsze reguły klasyfikacyjne.

Najbardziej ogólnie algorytm tworzenia drzewa klasyfikacyjnego można przedstawić następująco (Gatnar 1998):

- ◆ Dla danego zbioru obiektów sprawdzić, czy należą do tej samej klasy; jeśli tak – zakończyć postępowanie; jeśli nie - rozważyć wszystkie możliwe podziały danego zbioru na możliwie najbardziej jednorodne podzbiory.
- ◆ Ocenic jakość każdego z tych podzbiorów zgodnie z przyjętym kryterium i wybrać najlepszy z nich.
- ◆ Podzielić zbiór obiektów w wybrany sposób. Podział ten dokonywany jest w oparciu o charakterystykę obiektów (wartości cech), dlatego też cecha będąca podstawą podziału nie może być wybrana losowo. Do tego celu wykorzystywane są zatem różne miary opierające się



na założeniu, że istnieje związek między wartościami cech obiektów a ich przynależnością do określonej klasy.

- ♦ Wykonać powyższe kroki rekurencyjnie dla każdego z podzbiorów.

Cechami charakterystycznymi drzew klasyfikacyjnych są:

1. **Hierarchiczna natura** – polegająca na tym, że zależność liścia od drzewa, na którym rośnie, przez hierarchię podziałów gałęzi (począwszy od pnia) prowadzących do ostatniej gałęzi, z której wyrasta liść, można opisać w postaci pytań zadawanych w porządku hierarchicznym. Ostateczne podjęte decyzje zależą od odpowiedzi na wszystkie poprzednie pytania; wyraziście odróżnia je to od jednoczesnej natury decyzji powziętych w oparciu o np. analizę funkcji dyskryminacyjnych.
2. **Elastyczność** – polegająca na możliwości badania wpływu zmiennych predykcyjnych pojedynczo a nie wszystkich jednocześnie oraz na wykorzystywaniu zmiennych predykcyjnych różnych typów (przy podziałach jednowymiarowych można wykorzystywać predyktory nominalne, ciągłe lub jednocześnie obu typów) i osłabieniu założeń dotyczących ich pomiaru. Istnieje również możliwość klasyfikacji obiektów opisanych wektorem danych z występującymi wartościami brakującymi. W przypadku predyktorów ciągłych mierzonych co najmniej na skali interwałowej, istnieje możliwość definicji podziałów wykorzystujących kombinacje liniowe analogiczne do występujących w liniowej analizie dyskryminacji, jednakże podejście rekursywne przy drzewach klasyfikacyjnych nie ma ograniczenia polegającego na tym, liczba liniowych funkcji dyskryminacyjnych ma być równa mniejszej z dwóch liczb (liczba zmiennych predykcyjnych, liczba klas-1).

Metody tworzenia drzew klasyfikacyjnych

Do tworzenia drzew klasyfikacyjnych wykorzystać można moduł *Drzewa klasyfikacyjne* z pakietu *STATISTICA*, który stanowi pełną implementację programów *QUEST (Quick, Unbiased, Efficient Statistical Trees*, oprac. przez Loha i Shiha (1997)) i *CART (Classification And Regression Trees* oprac. przez Breimana i in w 1984 r.).

W procesie tworzenia drzewa klasyfikacyjnego wyróżnia się cztery etapy:

1. Określenie kryteriów trafności przewidywania.
2. Wybranie podziałów.
3. Wyznaczenie końca podziałów.
4. Wybranie drzewa właściwej wielkości.

Ad. 1: Za najbardziej trafne przewidywanie uważa się to, które wymaga najmniejszych kosztów. Koszty odnoszą się zwykle do proporcji błędnie zaklasyfikowanych przypadków, jednak zdarzają się także sytuacje, kiedy niektóre chybione przewidywania są o wiele bardziej katastrofalne w skutkach lub występują częściej niż inne, a wówczas problem proporcji błędnych zaklasyfikowań ustępuje zagadnieniu minimalizacji kosztów.

Na tym etapie pojawia się także problem zdefiniowania prawdopodobieństw *a priori* należenia obserwacji do klasy (określających, na ile jest prawdopodobne, bez żadnej wstępnej wiedzy na temat wartości zmiennych predykcyjnych w modelu, że dany przypadek znajdzie się w danej klasie). Równe prawdopodobieństwa *a priori* stosujemy, gdy w każdej z klas występuje mniej więcej równa liczba obserwacji. W przypadku próby losowej zróżnicowane stopy bazowe znajdują



odzwierciedlenie w wielkościach klas i tak też szacujemy prawdopodobieństwa *a priori*. Zdarzyć się także może, iż posiadamy pewną wiedzę zdobytą np. w trakcie wcześniejszych badań – wtedy prawdopodobieństwa *a priori* określamy zgodnie z tą wiedzą. Względne wielkości prawdopodobieństwa *a priori* należenia obserwacji do klas można wykorzystać do skorygowania ważności błędnych klasyfikacji dla każdej klasy.

Minimalizacja kosztów odpowiada minimalizacji ogólnej proporcji błędnych klasyfikacji wtedy, gdy prawdopodobieństwa *a priori* są proporcjonalne do wielkości klas a koszty błędnych klasyfikacji równe dla każdej klasy, co wynika z faktu, że lepsze przewidywanie w przypadku większych klas daje niższą ogólną stopę klasyfikacji błędnych.

W przypadku kosztów błędnych klasyfikacji należy również wziąć pod uwagę fakt, iż niekiedy pożądana jest bardziej trafna klasyfikacja w przypadku pewnych klas niż w przypadku innych. Może tak być m. in. w diagnostyce medycznej, kiedy bardziej pożądanym jest np. rozpoznanie pacjenta zagrożonego możliwością wystąpienia powikłań i zgonu. Niewiele tracimy, jeśli zaklasyfikujemy pacjenta do grupy wyższego ryzyka a jego rekonwalescencja przebiegnie bez problemów, a bardzo dużo – jeśli pacjenta zagrożonego możliwością wystąpienia powikłań błędnie zaklasyfikujemy do grupy o niskim ryzyku zgonu.

Określenie kosztów błędnych klasyfikacji można także wykorzystać do przeważenia analizy w kierunku nadreprezentowania niektórych klas w stosunku do innych.

Ad. 2: Podział wybierany jest w oparciu o zmienne wykorzystywane do przewidywania przynależności obiektów do klas wyznaczonych przez zmienną zależną. Podziały te są wybierane pojedynczo, od podziału przy węźle źródłowym, przez podziały wynikowych węzłów potomków aż do momentu, w którym dzielenie przerywa się a węzły nie podzielone stają się węzłami końcowymi (liśćmi).

Moduł *Drzewa klasyfikacyjne* daje możliwość wyboru jednej z trzech metod podziału:

- ◆ dyskryminacyjne podziały jednowymiarowe;
- ◆ dyskryminacyjne podziały z wykorzystaniem kombinacji liniowych;
- ◆ metoda CART wyczerpującego poszukiwania podziałów jednowymiarowych.

Szczegółowy opis procedur podaje Loh & Shih (1997), Breiman i in (1984), StatSoft, Inc (1995).

Ad. 3: Wyznaczenie końca podziałów, czyli określenie momentu zaprzestania podziałów, jest etapem szczególnie istotnym, bowiem brak takiego ograniczenia prowadzi do czystej, mało realistycznej klasyfikacji, w której każdy liść zawiera tylko jedną klasę obiektów. Dzielenie kontynuujemy do momentu, gdy wszystkie węzły końcowe są czyste lub zawierają nie więcej niż określoną minimalną liczbę obiektów bądź nie więcej niż określoną minimalną frakcję wielkości jednej lub więcej klas.

Ad. 4: Drzewo właściwej wielkości wyjaśnia fakty w sposób złożony, ale i na tyle prosty, na ile to możliwe; wykorzystuje te informacje, które zwiększają trafność przewidywań i pomija nieprzydatne; wreszcie - prowadzi do lepszego zrozumienia zjawisk, których dotyczy. Wyróżnia się dwie strategie wyboru drzewa właściwej wielkości spośród wszystkich możliwych drzew:

Pierwsza strategia polega na określeniu frakcji obiektów, która pozwala drzewu rozrastać się do pożądanej wielkości. Właściwa wielkość drzewa jest wyznaczana przez użytkownika na podstawie wiedzy z poprzednich badań, informacji diagnostycznych z wcześniejszych analiz czy intuicji badacza. Badanie sensowności wyboru wielkości drzewa odbywa się poprzez wykonanie tzw. sprawdzianu krzyżowego przeprowadzonego:



- ◆ na podstawie próby testowej zebranej niezależnie od uczącej lub wydzielonej z niej losowo;
- ◆ metodą *V-krotnego sprawdzianu krzyżowego* – gdzie z próby uczącej wyodrębnia się pewną liczbę podprób losowych zbliżonych do siebie wielkością. Drzewo oblicza się v razy opuszczając za każdym razem jedną z podprób i wykorzystując ją jako próbę testową. Koszty sprawdzianu krzyżowego oblicza się uśredniając koszty otrzymane dla każdej z v prób testowych.
- ◆ metodą *globalnego sprawdzianu krzyżowego* – w którym całą analizę powtarza się określoną liczbę razy eliminując część próby uczącej równą 1 dzielone przez określoną liczbę; każda wyeliminowana próba jest następnie wykorzystana jako próba testowa w sprawdzianie krzyżowym wybranego drzewa klasyfikacyjnego.

Drugą strategią wyboru właściwej wielkości drzewa jest przycinanie na podstawie minimalizacji kosztów i złożoności drzewa w sprawdzianie krzyżowym. Wykorzystuje się tutaj procedury opracowane przez Breimana i in. (1984). Zamiast określania kryterium zaniechania podziału zbioru obiektów tworzy się pełne drzewo klasyfikacyjne, które następnie jest porządkowane poprzez zamianę węzłów wewnętrznych na liście tak, aby przy zmniejszaniu się jednorodności klas nie wzrastał błąd klasyfikacji. Szczegółowe techniki postępowania przedstawia StatSoft, Inc. (1995) oraz Gatnar (1998).

Wykorzystanie drzew klasyfikacyjnych

Material badawczy

Choroba wieńcowa stanowi poważny problem dla opieki zdrowotnej nie tylko w Polsce, ale i na świecie, dotyczy bowiem wszystkich społeczeństw wysokocywilizowanych i powszechnie uważana jest za epidemię XX wieku. Zakwalifikowanie pacjenta z chorobą wieńcową do leczenia operacyjnego jest przykładem decyzji podejmowanej w warunkach niepewności. W takim przypadku niezbędne staje się porównanie korzyści z operacji (przedłużenie życia, zmniejszenie prawdopodobieństwa nagłego zgonu) z ryzykiem wystąpienia powikłań i zgonu. Przy szacowaniu tego ryzyka należy jednocześnie rozważyć szereg informacji dotyczących charakterystyki pacjenta, przebiegu choroby, wyników badań itd.

W celu zwiększenia bezpieczeństwa operacji wskazane jest opracowanie procedury pozwalającej ustalać poziom ryzyka operacyjnego poprzez klasyfikację pacjenta do odpowiedniej grupy wyznaczającej określony stopień tegoż ryzyka.

Badaniu podlegają pacjenci Kliniki Kardiochirurgii, którzy w latach 1997-1999 zostali poddani operacji w związku z chorobą niedokrwinną serca. Każdy pacjent jest opisany wektorem cech określających jego stan przed i w trakcie operacji oraz przebieg leczenia około i pooperacyjnego. Klasyfikacja pacjentów do wyodrębnionych grup ryzyka i określenie czynników ryzyka pozwala na wprowadzenie pewnej zgodności postępowania różnych osób (np. lekarzy różnych specjalności) stojących przed problemem klasyfikacji operowanych poprzez zwrócenie uwagi lekarza na ewentualne komplikacje i zagrożenie życia pacjenta.

Pacjentów podzielono na dwie grupy: (1) – osoby, u których nie pojawiły się powikłania kardiologiczne w trakcie i po operacji, a ich rekonwalescencja przebiegła pomyślnie oraz (2) – osoby, u których nastąpił zgon w trakcie operacji lub w czasie pobytu na OIOM.



Zmienną zależną jest zatem zmienna zerojedynkowa ZGON, przyjmująca wartość 0 dla osób, u których operacja zakończyła się sukcesem oraz wartość 1 – dla osób zmarłych.

Zmienne określające czynniki ryzyka wyodrębniono badając korelacje między daną zmienną a zmienną opisującą wystąpienie zgonu, na podstawie już istniejących kart ryzyka i wcześniejszych badań innych autorów oraz na podstawie informacji lekarzy z ich codziennej praktyki. Lista zmiennych predykcyjnych przedstawia się następująco:

1. CAD – obciążenie rodzinne chorobą wieńcową (0 – nie, 1 – tak);
2. Cukrzyca (0 – nie, 1 – tak);
3. AO – miażdżyca zarostowa kończyn dolnych (0 – nie, 1 – tak);
4. Nadczynność tarczycy (0 – nie, 1 – tak);
5. Poprzednie interwencje kardiochirurgiczne (0 – nie, 1 – tak);
6. Zwężenie pnia powyżej 75% (0 – nie, 1 – tak);
7. Tryb zabiegu (planowy, pilny, nagły);
8. Wiek w latach;
9. BSA – wskaźnik masy ciała liczony wg formuły $\sqrt{\frac{\text{waga(kg)} * \text{wzrost(cm)}}{3600}}$;
10. RRS – ciśnienie skurczowe w mm Hg;
11. RRD – ciśnienie rozkurczowe w mm Hg;
12. EF% - frakcja wyrzutowa lewej komory serca;
13. GOT – próba wątrobowa.

Zebrane dane dotyczyły 352 pacjentów. Dane z lat 1997-1998 (192 osoby, w tym 46 zgonów) wykorzystano jako próbę uczącą, a z roku 1999 (160 osób, w tym 11 zgonów) – jako próbę testową.

Wyniki

W przykładzie 1 do stworzenia drzewa klasyfikacyjnego wybrano metodę dyskryminacyjnych podziałów jednowymiarowych dla predyktorów nominalnych i porządkowych.

Z badań ośrodków medycznych i własnych doświadczeń lekarzy wynika, że ryzyko zgonu pacjenta operowanego z powodu choroby niedokrwiennej serca wynosi ok. 10%, dlatego też prawdopodobieństwa *a priori* należenia obserwacji do klas ustalono odpowiednio: 0,90 (dla klasy „bez zgonu”) i 0,10 (dla klasy „zgon”).

Ze względu na fakt, iż bardziej pożądana jest trafna klasyfikacja pacjentów zagrożonych zgonem, a także z powodu znacznie mniejszej liczebności klasy „zgon” przyjęto, że koszt błędnej klasyfikacji pacjenta z klasy „zgon” jako pacjenta niezagrażonego jest czterokrotnie wyższy niż koszt klasyfikacji pacjenta niezagrażonego do grupy wysokiego ryzyka.

Jako regułę stopu przyjęto „bezpośrednie zatrzymanie typu FACT” przy frakcji obiektów równej 0,15.

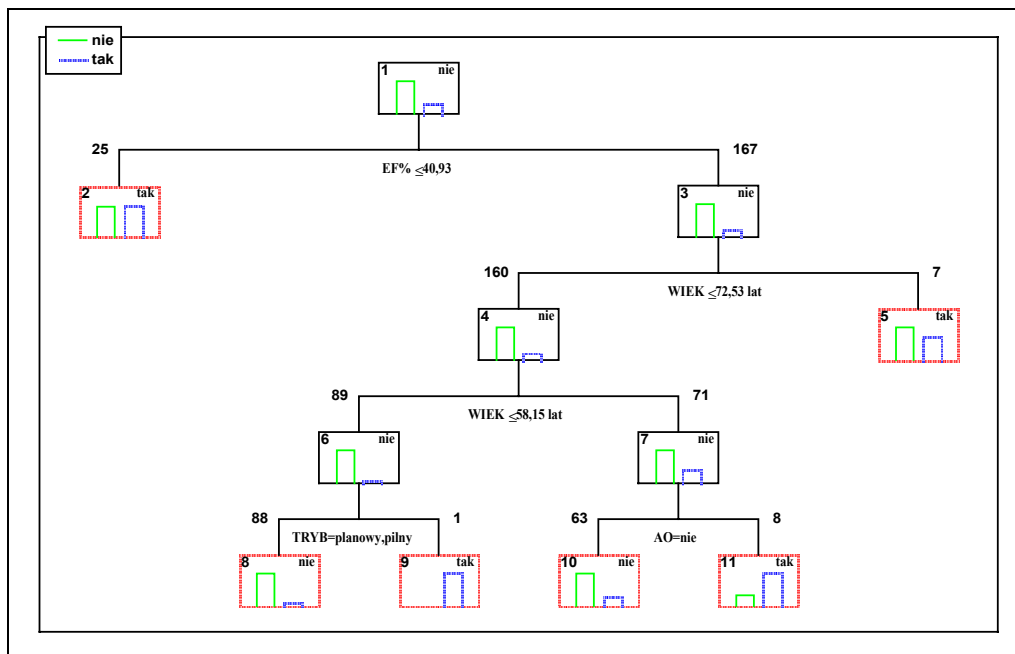
Po uzyskaniu podziałów jednowymiarowych zmienne predykcyjne można porangować w skali od 0 do 100, w zależności na ile ważny jest ich wpływ na wartości zmiennej zależnej. Wyniki przedstawia tabela 1. Jak widać, największy wpływ na zmienną „zgon” ma wiek pacjenta. Jest to zgodne z informacjami lekarzy z ich codziennej praktyki – pacjenci wysokiego ryzyka są starsi, więcej jest również osób po 70 roku życia.

Tabela 1. Ranking ważności predyktorów

| Zmienna | Ranking |
|---|---------|
| Wiek | 100 |
| Poprzednie interwencje kardiochirurgiczne | 77 |
| AO | 65 |
| EF% | 65 |
| Tryb zabiegu | 64 |
| Nadczynność tarczycy | 57 |
| Zwężenie pnia | 25 |
| RRD | 25 |
| CAD | 21 |
| BSA | 21 |
| RRS | 15 |
| GOT | 11 |
| Cukrzyca | 5 |

Źródło: Obliczenia własne

Otrzymane drzewo przedstawia rys. 1, a jego strukturę opisuje tabela 2. Drzewo obejmuje 5 podziałów i ma 6 węzłów końcowych.



Rys. 1. Drzewo klasyfikacyjne (dyskryminacyjne podziały jednowymiarowe)

Tabela 2. Struktura drzewa

| Nr węzła | Gałąź | | Liczba przypadków w węźle | | Przewidywana klasa | Podział | | |
|----------|--------------|---------------|---------------------------|--------------|--------------------|---------|---------|-----------|
| | Lewo-stronna | Prawo-stronna | klasa "bez zgonu" | klasa "zgon" | | stała | zmienna | kategoria |
| 1 | 2 | 3 | 146 | 46 | nie | -40,93 | EF% | |
| 2 | | | 12 | 13 | tak | | | |
| 3 | 4 | 5 | 134 | 33 | nie | -72,528 | WIEK | |
| 4 | 6 | 7 | 130 | 30 | nie | -58,15 | WIEK | |
| 5 | | | 4 | 3 | tak | | | |
| 6 | 8 | 9 | 80 | 9 | nie | | TRYB | planowy |
| 7 | 10 | 11 | 50 | 21 | nie | | AO | nie |
| 8 | | | 80 | 8 | nie | | | |
| 9 | | | 0 | 1 | tak | | | |
| 10 | | | 48 | 15 | nie | | | |
| 11 | | | 2 | 6 | tak | | | |

Źródło: Obliczenia własne

Na początku wszystkie 192 osoby przypisano do węzła źródłowego i tymczasowo zaklasyfikowano do grupy „bez zgonu” (na rys. 1 – „bez zgonu” = „nie”, „zgon” = „tak”). Węzeł źródłowy dzieli się na dwa nowe węzły. Tekst poniżej węzła źródłowego opisuje ten podział. Pacjenci, u których frakcja wyrzutowa lewej komory serca nie przekracza 40,9% (25 osób) zostają przypisani do węzła nr 2 i zaklasyfikowani do grupy „zgon” a pozostali (167 osób) – do węzła nr 3 i zaklasyfikowani do grupy „bez zgonu”. Węzeł nr 2 jest węzłem końcowym i nie ulega już podziałowi.

Węzeł nr 3 dzieli się wg kryterium $WIEK \leq 72,5$ lat. Pacjenci w wieku co najwyżej 72,5 lat zostają przypisani do węzła nr 4 i zaklasyfikowani do grupy „bez zgonu”, a pozostali – do węzła nr 5 – grupa „zgon”. Węzeł nr 5 jest węzłem końcowym.

Węzeł nr 4 dzieli się wg kryterium $WIEK \leq 58,2$ lat. Pacjenci w wieku do 58,2 lat zostają przypisani do węzła nr 6, który dzieli się dalej wg zmiennej „Tryb zabiegu”. Osoby, które operowano w trybie planowym lub pilnym zostają przypisane do węzła nr 8 i zaklasyfikowane do grupy niezagrożonej zgonem, a osoby operowane w trybie nagłym – przypisano do węzła nr 9 jako zagrożone zgonem.

Chorzy w wieku powyżej 58,2 lat zostali przypisani do węzła nr 7, który dzieli się dalej zgodnie z kryterium $AO = \text{nie}$. Pacjenci, u których nie występowała miażdżycza zarostowa kończyn dolnych zostali zaklasyfikowani do grupy „bez zgonu” (węzeł nr 10) a pozostali – do grupy zagrożonej zgonem (węzeł nr 11).

Reasumując, pacjentów zagrożonych zgonem można opisać jako osoby:

- ◆ O frakcji wyrzutowej lewej komory serca poniżej 40,9%;
- ◆ W wieku powyżej 72,5 lat;
- ◆ W wieku do 58,2 lat i operowane w trybie nagłym;
- ◆ W wieku powyżej 58,2 lat chorujące na miażdżycę zarostową kończyn dolnych.



Histogramy wykreślone wewnątrz końcowych węzłów pokazują, że uzyskana klasyfikacja nie jest bezbłędna. W tabeli 3 przedstawione są odsetki błędnych klasyfikacji dla próby uczącej i dla próby testowej.

Tabela 3. Odsetek błędnych klasyfikacji

| Klasa przewidywana | Próba ucząca Koszt sprawdzianu krzyżowego=0,601 | | | Próba testowa Koszt sprawdzianu krzyżowego=0,247 | | |
|--------------------|--|------|-------------------------------|---|------|-------------------------------|
| | Klasa obserwowana | | Odsetek błędnych klasyfikacji | Klasa obserwowana | | Odsetek błędnych klasyfikacji |
| | bez zgonu | zgon | | bez zgonu | zgon | |
| bez zgonu | 128 | 23 | 50,00% | 119 | 3 | 27,27% |
| zgon | 18 | 23 | 12,33% | 30 | 8 | 20,13% |
| razem | 146 | 46 | 21,35% | 149 | 11 | 20,63% |

Źródło: Obliczenia własne

Analiza tabeli 3 wskazuje, iż w próbie testowej lepsze klasyfikacje otrzymujemy w przypadku grupy pacjentów niezagrażonych zgonem, jednak właściwym sprawdzianem przydatności drzewa są odsetki błędnych klasyfikacji dla próby testowej, której nie wykorzystano do tworzenia drzewa. Jak widać, poprawie uległy klasyfikacje pacjentów z grupy zagrożonej zgonem – błędnie zaklasyfikowano jedynie 27,3% badanych. Ogólny odsetek nieprawidłowych klasyfikacji oraz koszt sprawdzianu krzyżowego także są mniejsze dla próby testowej.

W przykładzie 2 do utworzenia drzewa klasyfikacyjnego wykorzystano metodę CART wyczerpującego poszukiwania podziałów jednowymiarowych. Prawdopodobieństwa *a priori* oraz koszty błędnych klasyfikacji pozostają bez zmian. Jako regułę stopu wybrano „przycięcie przy błędzie złej klasyfikacji” z minimalną liczebnością równą 20 przypadków.

Analiza rankingu predyktorów w tym przypadku wskazuje, iż największy wpływ na zmienną zależną ma frakcja wyrzutowa lewej komory serca, a w następnej kolejności wskaźnik powierzchni ciała oraz próba wątrobowa GOT.

W tabeli 4 przedstawiono sposób wyboru drzewa klasyfikacyjnego. Wybrane drzewo ma 6 węzłów końcowych, koszt sprawdzianu krzyżowego równy 0,465 z błędem standardowym 0,0356, koszt resubstytucji równy 0,190 i wygładzoną wartość złożoności węzła równą 0,0011.

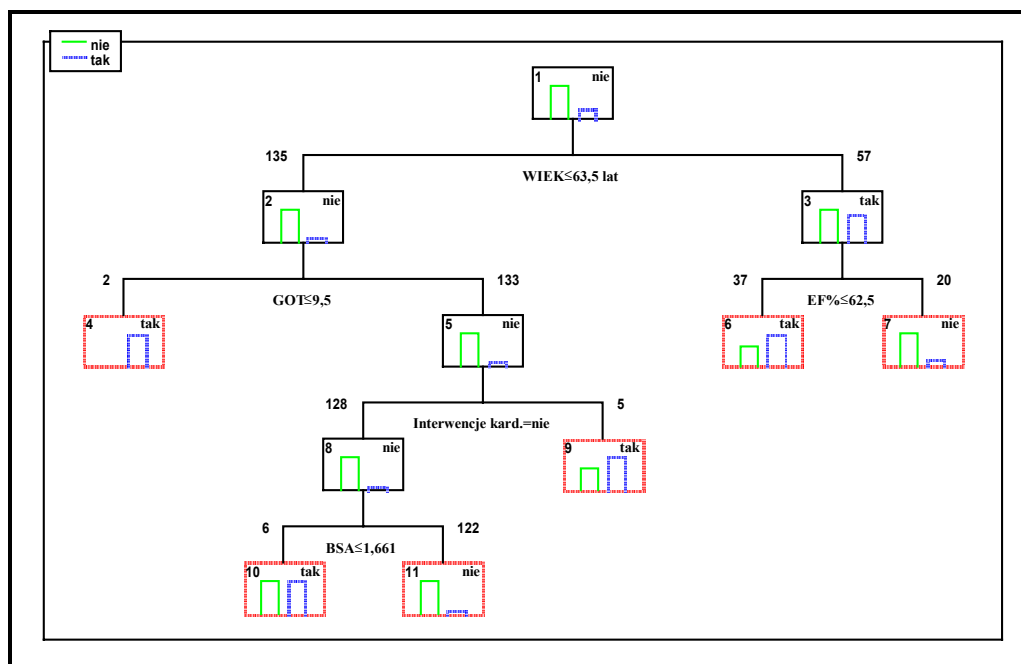
Wybór drzewa właściwej wielkości odbywa się poprzez analizę kosztów sprawdzianu krzyżowego. Wybiera się to drzewo o najmniejszej wielkości, którego koszty sprawdzianu krzyżowego nie różnią się znacznie od minimalnych kosztów sprawdzianu krzyżowego (czyli którego koszty sprawdzianu krzyżowego nie przekraczają minimalnych kosztów plus 1 błąd standardowy dla drzewa o minimalnych kosztach sprawdzianu krzyżowego).

Tabela 4. Wybór drzewa klasyfikacyjnego

| Nr drzewa | Liczba węzłów końcowych | Koszt spr. Krzyżowego | Błąd std. | Koszt resubstytucji | Węzeł złożoności |
|-----------|-------------------------|-----------------------|-----------|---------------------|------------------|
| 1 | 12 | 0,4571 | 0,0362 | 0,1854 | 0 |
| 2 | 7 | 0,4571 | 0,0362 | 0,1893 | 0,0008 |
| *3 | 6 | 0,4655 | 0,0356 | 0,1904 | 0,0011 |
| 4 | 5 | 0,5313 | 0,0294 | 0,1963 | 0,0058 |
| 5 | 4 | 0,6081 | 0,0234 | 0,2069 | 0,0106 |
| 6 | 3 | 0,5958 | 0,0254 | 0,2202 | 0,0134 |
| 7 | 1 | 0,6923 | 0 | 0,3077 | 0,0437 |

Źródło: Obliczenia własne

Drzewo klasyfikacyjne przedstawia rys. 2.



Rys. 2. Drzewo klasyfikacyjne (metoda CART)

W tym przypadku, pacjentów zagrożonych możliwością wystąpienia powikłań i zgonu możemy scharakteryzować jako osoby:

- ◆ W wieku powyżej 63,5 lat;
- ◆ O wielkości GOT poniżej 9,5 jednostek;
- ◆ Które poddane były wcześniej operacjom serca;
- ◆ O powierzchni ciała poniżej 1,66 m².

Tabela 5 przedstawia odsetki błędnych klasyfikacji w próbie uczącej i próbie testowej.



Tabela 5. Odsetek błędnych klasyfikacji

| Klasa przewidywana | Próba ucząca Koszt sprawdzianu krzyżowego=0,465 | | | Próba testowa Koszt sprawdzianu krzyżowego=0,502 | | |
|--------------------|--|------|-------------------------------|---|------|-------------------------------|
| | Klasa obserwowana | | Odsetek błędnych klasyfikacji | Klasa obserwowana | | Odsetek błędnych klasyfikacji |
| | bez zgonu | zgon | | bez zgonu | zgon | |
| bez zgonu | 127 | 15 | 32,61% | 110 | 7 | 63,64% |
| zgon | 19 | 31 | 13,01% | 39 | 4 | 26,17% |
| razem | 146 | 46 | 17,71% | 149 | 11 | 28,75% |

Źródło: Obliczenia własne

Jak łatwo zauważyć, lepsze klasyfikacje uzyskaliśmy dla próby uczącej. Algorytm CART gwarantuje znalezienie podziałów, które stanowią najlepszą klasyfikację w próbie uczącej, ale niekoniecznie w próbach sprawdzania krzyżowego.

Uwagi końcowe

- ◆ Drzewa klasyfikacyjne mogą być alternatywą dla metod tradycyjnych, wówczas gdy nie ma możliwości spełnienia założeń wymaganych przez te metody.
- ◆ Wykres drzewa przedstawia wszystkie informacje w sposób prosty, bezpośredni i łatwy do interpretacji.
- ◆ Duża elastyczność drzew klasyfikacyjnych pozwala na badanie wpływu poszczególnych zmiennych na zmienną objaśnianą pojedynczo a nie łącznie.
- ◆ Wykorzystanie drzew klasyfikacyjnych ułatwia również analizę danych z wartościami brakującymi, które można zastąpić odpowiednią zmienną tekstową i potraktować jako szczególny przypadek predyktora nominalnego.

Literatura

1. Breiman L. i in. (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
2. Ferraris V. i in. (1996), *Risk Factors for Postoperative Morbidity*, The Journal of Thoracic and Cardiovascular Surgery, s. 731-741.
3. Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.
4. Loh W.-Y., Shih Y.-S. (1997), *Split Selection Methods for Classification Trees*, Statistica Sinica, 7, s. 815-840.
5. Spivack S. D. i in. (1996), *Preoperative Prediction of Postoperative Respiratory Outcome*, CHEST, s. 1222-1230.
6. StatSoft, Inc. (1995), *STATISTICA for Windows* [Computer program manual], Tulsa.