PRACA POGLĄDOWA

Drzewa klasyfikacyjne w medycynie

Classification trees in medicine

Aleksander J. Owczarek

Received: 12.12.2013
Revised: 19.01.2014
Accepted: 20.01.2014
Published online: 31.12.2014

STRESZCZENIE

W pracy zaprezentowano wykorzystanie w medycynie komputerowych systemów diagnostyki medycznej. Przedstawiono budowę klasycznego drzewa decyzyjnego oraz zalety i wady stosowania drzew klasyfikacyjnych. Ponadto omówiono działanie drzew klasyfikacyjnych w świetle innych klasycznych metod statystycznych, takich jak analiza dyskryminacyjna czy regresja logistyczna, z uwzględnieniem problemu współliniowości zmiennych czy problemu występowania tzw. danych niepełnych. Podano wybrane przykłady zastosowania drzew klasyfikacyjnych w medycynie.

SŁOWA KLUCZOWE
drzewa klasyfikacyjne, proces decyzyjny, współliniowość zmiennych, dane niepełne

ABSTRACT

The paper presents the use of computerized diagnostic decision support systems for medical diagnostics in medicine. The structure of a classical decision tree and the advantages and disadvantages of using classification trees have been discussed. Moreover, the paper deals with the effect of classification trees with respect to other classic statistical methods, such as discriminant analysis and logistic regression, taking into account the problem of variable multicollinearity and the problem of the occurrence of so-called missing data. Additionally, some examples of the application of classification trees in medicine have been shown.

KEY WORDS
classification trees, decision process, multicollinearity, missing data

Zakład Statystyki Katedry Analizy Instrumentalnej
Wydziału Farmaceutycznego
z Oddziałem Medycyny Laboratoryjnej w Sosnowcu
Śląskiego Uniwersytetu Medycznego
w Katowicach

ADRES DO KORESPONDENCJI:

Dr hab. n. o zdr. inż. Aleksander Jerzy Owczarek
Zakład Statystyki Katedry Analizy Instrumentalnej
Wydziału Farmaceutycznego z Oddziałem Medycyny
Laboratoryjnej w Sosnowcu
Śląskiego Uniwersytetu Medycznego
w Katowicach
ul. Ostrogórska 30
41-200 Sosnowiec
tel. + 48 32 364 13 28; + 48 664 945 174
e-mail: aowczarek@sum.edu.pl

Ann. Acad. Med. Siles. 2014, 68, 6, 449–456

eISSN 1734-025X
www.annales.sum.edu.pl

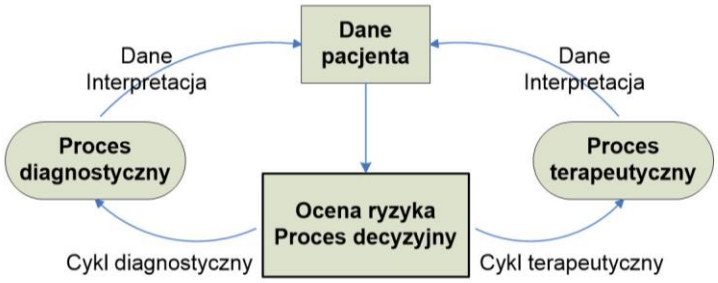
1. Komputerowe systemy wspomagania diagnostyki medycznej (KSWDM)

Komputerowe systemy diagnostyki medycznej (KSWDM), wykorzystujące różnego rodzaju metodologię wnioskowania, są powszechnie stosowane w wielu dziedzinach medycyny, co zostało dobrze udokumentowane [1,2,3,4,5,6,7]. W procesie ich projektowania i realizacji wykorzystuje się nie tylko klasyczne metody statystyczne (wielowymiarową regresję logistyczną, analizę dyskryminacyjną, klasyfikatory Bayesa czy metodę *k*-najbliższych sąsiadów), lecz także eksplorację danych oraz sztuczną inteligencję (w tym sieci neuronowe, logikę rozmytą, sieci bayesowskie, maszyny wektorów nośnych, drzewa klasyfikacyjne i regresyjne) [8,9,10,11,12,13,14,15, 16,17,18]. Ich przeznaczenie obejmuje m.in. wspomaganie prowadzenia badań przesiewowych, procesów diagnostycznego (z uwzględnieniem procedur laboratoryjnych) oraz terapeutycznego (w tym dawkowania leków i farmakoekonomiki), a także zarządzania systemem opieki zdrowotnej w chorobach przewlekłych [3,4,19,20]. Szacuje się, że zastosowanie KSWDM w około 60–70% przeanalizowanych przypadków istotnie poprawiło jakość opieki zdrowotnej w praktyce klinicznej, w około 60% systemów zarządzania w chorobach przewlekłych, w ponad 65% jakość systemów farmaceutycznych oraz znacząco poprawiło satysfakcję pacjenta. Istotnym problemem – zarówno z klinicznego, jak i ekonomicznego punktu widzenia – – jest ocena ryzyka niekorzystnych zdarzeń, takich jak hospitalizacja lub zgon. Wykorzystanie KSWDM zmniejszyło ryzyko postawienia błędnej diagnozy średnio o 16% [1,6,21,22,23,24,26].

W większości obecnie istniejących systemów ochrony zdrowia podejmowanie decyzji dotyczących diagnostyki i terapii jest u każdego pacjenta procesem złożonym i skomplikowanym [3,20]. Zaprezentowany na rycinie 1 proces decyzyjny nieodłącznie wiąże się z ograniczeniami narzucanymi przez system opieki zdrowotnej, warunkami klinicznymi, dostępnymi informacjami, preferencjami pacjenta, personelu medycznego i zarządzającego daną placówką medyczną [20,27]. Warunki kliniczne obejmują z jednej strony naturę, nasilenie i złożoność analizowanego problemu (przypadku), z drugiej zaś politykę zdrowotną i rygory ekonomiczne obowiązujące lekarza w danej instytucji. Preferencje pacjenta są najistotniejsze, gdy w procesie decyzyjnym brak wyraźnie zaznaczonego dalszego kierunku działań (istnieje wiele możliwości decyzji co do sposobu diagnostyki i leczenia). Trzeba wówczas uwzględnić satysfakcję pacjenta i jego dalszą jakość życia, wynikającą z podjętej decyzji. Ostateczna decyzja wymaga nie tylko ekonomicznej analizy procedur diagnostycznych i terapeutycznych, ale także sytuacji społeczno-ekonomicznej pacjenta [28,29, 30,31].

Istotnym aspektem systemu opieki zdrowotnej są błędy pojawiające się w procesie decyzyjnym, wynikające z jednej strony z braku informacji dotyczących konkretnego pacjenta (sytuacji) w danej chwili czasowej i środowisku pracy, z drugiej zaś z niepewności związanej z aktualnie dostępnymi danymi (błędy pomiarowe, wyniki fałszywie dodatnie oraz fałszywie ujemne). Aby zminimalizować ryzyko podjęcia nieprawidłowej decyzji, w literaturze oraz w systemach zarządzania w medycynie, proponuje się nowy paradygmat opieki zdrowotnej, postulujący tworzenie multidyscyplinarnych zespołów złożonych z lekarzy klinicystów, personelu medycznego oraz biostatystyków i/lub inżynierów biomedycznych.

Głównym celem działania takich zespołów jest optymalizacja procesu decyzyjnego, uwzględniająca zarówno warunki kliniczne, jak i satysfakcję oraz jakość życia pacjenta. Efektem pracy zespołu ma być poprawa funkcjonowania systemu opieki zdrowotnej i odległego rokowania chorych, uzyskiwana na podstawie



Ryc. 1. Schemat procesu decyzyjnego (oceny ryzyka) obejmujący informacje związane z pacjentem oraz cykle diagnostyczny i terapeutyczny.
Fig. 1. Diagram of decision-making process (risk assessment), including information related to patient, with diagnostic and therapeutic cycles.

opracowanych schematów postępowania i komputerowych systemów wspomagania diagnostyki medycznej [3,6,21]. Nie bez znaczenia jest również to, że w wielu systemach opieki zdrowotnej wypracowano efektywne metody zarządzania w sytuacjach kryzysowych (stanach ostrych), słabiej jednak poradzono sobie z kontrolą w przypadku schorzeń przewlekłych [32].

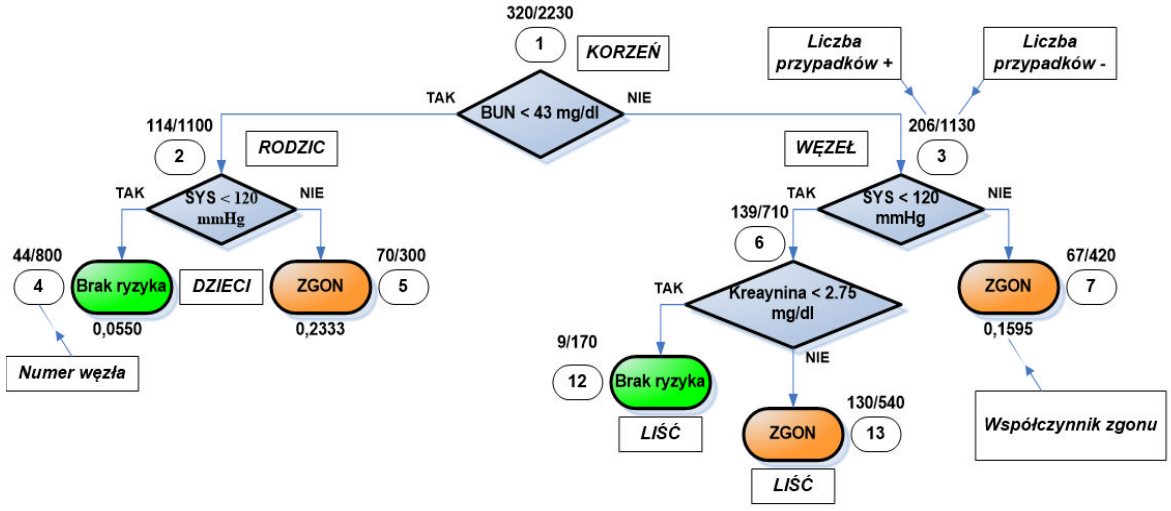
2. Drzewa klasyfikacyjne

Drzewa klasyfikacyjne (zwane też decyzyjnymi) stanowią rodzinę metod statystycznych, dokonującą za pomocą diagramów (tzw. nieskierowanych acyklicznych grafów spójnych) sekwencyjnego podziału badanej przestrzeni danych na klasy (podprzestrzenie) o podobnych właściwościach. Historia drzew decyzyjnych rozpoczęła się wraz z ukazaniem się książki Breimana i wsp., prezentującej stosowany do dziś model CART (*Classification and Regression Tree*). Kolejną istotną pozycją była książka Quinlana omawiającą budowę i implementację tzw. algorytmu C4, będącego modyfikacją algorytmu zaproponowanego przez Breimana [33,34,35,36].

2.1. Budowa drzewa klasyfikacyjnego

Drzewo decyzyjne składa się z **korzenia** (w którym rozpoczyna się proces rekurencyjnego podziału) oraz **gałęzi** prowadzących od korzenia do kolejnych **węzłów**. Węzeł, z którego wychodzą gałęzie skierowane do kolejnych węzłów, nazywamy **rodzicem** węzłów, do których prowadzą gałęzie. **Potomkami** nazywamy wtedy węzły połączone z rodzicem. Węzeł nieposiadający dzieci (czyli taki, w którym nie następuje podział podprzestrzeni danych) nazywamy **liściem**. W liściu zawarta jest informacja o przynależności danych w podprzestrzeni do konkretnej klasy. W każdym węźle sprawdzany jest pewien warunek (zależny od typu drzewa), dotyczący danej obserwacji i na jego podstawie jest wybierana jedna z gałęzi prowadząca do kolejnego, położonego niżej węzła. Istotą tworzenia drzewa klasyfikacyjnego jest uzyskanie narzędzia (modelu) pozwalającego na klasyfikację przyszłych obserwacji, dla których nie posiadamy informacji o przynależności do konkretnej klasy. Utworzone drzewo jest modelem nie tylko predykcyjnym ale, co istotne w praktyce klinicznej, deskryptywnym, pozwalającym opisać i zaprezentować wzorce w badanej zbiorowości [37,38]. Proces budowy drzewa odbywa się metodą rekurencyjnego podziału [39], gdzie w każdym kolejnym węźle do analizy można wykorzystać inną zmienną niezależną. Na każdym bowiem etapie analizuje się wszystkie predykatory (zmienne) i wybiera ten, który zapewnia najlepszy podział węzła, czyli pozwala uzyskać najbardziej homogeniczne podzbiory [40,41,42]. Przykładową strukturę drzewa decyzyjnego zaprezentowano na rycinie 2.

Podział w danym węźle odbywa się tylko na podstawie wektorów próby uczącej, które dotarły do danego węzła i polega na najlepszym (w określonym sensie)



Ryc. 2. Przykładowe drzewo decyzyjne (klasyfikacyjne); BUN – stężenie mocznika w surowicy, SYS – skurczowe ciśnienie tętnicze.
Fig. 2. Example of decision tree (classification); BUN – blood urea nitrogen, SYS – systolic blood pressure.

rozdzieleniu tej podpróby na 2 części[1] (w przypadku drzew binarnych), przechodzące do potomków [43]. Podział ten powinien być taki, aby różnorodność otrzymanych części zbioru danych docierających do potomków była możliwie najmniejsza. Przeprowadzenie podziału wymaga m.in. podania stosownej miary różnorodności klas w węźle. Miary różnorodności stosowane w procesie tworzenia drzew decyzyjnych zostały obszernie omówione w pracy Loh i wsp. [42]. Typowe miary stosowane dla drzew binarnych to:

– indeks (wskaźnik) Giniego:

$$Q = 2 \cdot p \cdot (1 - p) \quad (1)$$

– entropia (wskaźnik entropii):

$$Q = -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p) \quad (2)$$

gdzie p oznacza prawdopodobieństwo przynależności do jednej z klas (zwykle reprezentującą analizowane zdarzenie, np. wystąpienie zgonu czy hospitalizacji).

2.2. Zalety i wady drzew decyzyjnych

Drzewa decyzyjne (DT) mają wiele zalet, m.in. [38,39,40,41,42,43,44]:

– przynależność do grupy metod nieparametrycznych (nie jest wymagana normalność rozkładu analizowanych zmiennych),
– niezależność procesu uczenia drzewa od zastosowanych wcześniej transformacji danych niezależnych (logarytmowanie, pierwiastkowanie itp. nie zmienia struktury drzewa, choć zmienia wartości punktów odcięcia – *splitting values*),
– identyfikację w procesie uczenia danych odstających (*outliers*), które trafiają następnie do osobnego węzła/liścia, nie wpływając na dalszy podział,
– szybkość obliczeniową,
– możliwość analizy danych zarówno numerycznych, jak i w skali porządkowej,
– łatwość zrozumienia i wizualizacji – istnieje możliwość utworzenia na ich podstawie reguł decyzyjnych,
– możliwość wykrywania złożonych interakcji między analizowanymi zmiennymi, które mogą nie zostać odkryte przy zastosowaniu klasycznych metod analizy wielowymiarowej,
– niewrażliwość na wystąpienie kolinearności oraz heteroskedastyczności wśród zmiennych.

Algorytmy oparte na drzewach decyzyjnych nie są, oczywiście, pozbawione wad. Wiarygodność procesu decyzyjnego realizowanego na podstawie drzewa decyzyjnego zależy od jakości danych (zbioru uczącego) użytych w procesie konstruowania drzewa. Drzewa decyzyjne mogą być wrażliwe na zmiany struktury zbioru uczącego – nawet niewielka zmiana danych wejściowych może skutkować znaczącą zmianą struktury drzewa decyzyjnego, obejmującą także miejsca i wartości podziału.

Do głównych wad drzew decyzyjnych należy ich złożona budowa. W przypadku dużych zbiorów danych ich wielogałęziowa konstrukcja przestaje być czytelna i zrozumiała, a sam proces tworzenia staje się bardzo czasochłonny. Ponadto algorytmy te nie są odporne na tzw. przekleństwo wielowymiarowości. Podjęcie decyzji na podstawie drzewa decyzyjnego z wieloma gałęziami i podziałami może być w efekcie znacznie spowolnione [41]. Budowa drzewa decyzyjnego realizowana jest z wykorzystaniem tzw. strategii „dziel i rządź" (*divide and conquer*), która sprawdza się w przypadku istnienia kilku istotnych zmiennych (atrybutów), natomiast działa mniej sprawnie w obecności wielu złożonych interakcji. Oprócz tego uzyskany model niekoniecznie będzie najlepszy w sensie optymalizacji globalnej, tzn. mimo istnienia pewnego optimum globalnego w przestrzeni rozpatrywanych zmiennych, utworzone drzewo decyzyjne przedstawi rozwiązanie będące optimum lokalnym. Istotnym aspektem jest również możliwość wystąpienia, dla niektórych modeli drzew decyzyjnych, obciążoności procesu wyboru zmiennych do podziału. W efekcie algorytm jest skłonny wybierać zmienne prowadzące do większej liczby podziałów (częstszy wybór zmiennych jakościowych w stosunku do zmiennych ilościowych lub wybór tych zmiennych ilościowych, które mają większa liczbę wartości) [41,45,46].

Klasyczne drzewa typu CART umożliwiają wyłącznie podział binarny. W celu podziału danych mających kilka podklas istnieje ryzyko utworzenia dużego drzewa o wielu poziomach, tak by zapewnić poprawność klasyfikacji. Istotną wadą drzew decyzyjnych typu CART jest to, że nie bazują one na żadnym modelu probabilistycznym. Nie istnieje zatem poziom istotności statystycznej czy przedział ufności związany z klasyfikacją nowych przypadków opartą na strukturze stworzonego drzewa. Rozwiązaniem tego problemu są drzewa umożliwiające podziały na wiele klas (typu CHAID) bądź realizujące bardziej zaawansowane metody podziału w węzłach (drzewa typu QUEST, CRUISE, LOTUS czy drzewa prawdopodobieństwa warunkowego) [42,44,45,46].

Należy również podkreślić, że realizacja algorytmów opartych na drzewach decyzyjnych – bez odpowiedniej wiedzy, doświadczenia i zaawansowania w zakresie analiz statystycznych – może prowadzić do uzyskania błędnych modeli i skutkować podjęciem nieprawidłowych decyzji [38,39,40,41,42,43,44,45,46].

[1]Istnieją drzewa decyzyjne umożliwiające podział grupy na więcej części (np. CHAID).

3. Drzewa decyzyjne a inne metody statystyczne

Jedną z powszechnie stosowanych metod statystycznych, pozwalającą wyznaczyć istotne atrybuty umożliwiające klasyfikację pacjentów, jest analiza dyskryminacyjna. Jej zastosowanie w analizie danych wymaga weryfikacji trzech najistotniejszych założeń: zapewnienia wielowymiarowej normalności analizowanych zmiennych, jednorodności wariancji/kowariancji oraz braku korelacji między średnimi a wariancjami [36]. Alternatywą analizy dyskryminacyjnej, często wykorzystywaną w medycynie, jest regresja logistyczna – model statystyczny umożliwiający opisanie wpływu kilku zmiennych zależnych (ilościowych i jakościowych) na dychotomiczną zmienną zależną. Stosowanie regresji logistycznej wymaga zdefiniowania *a priori* istotnych zmiennych wpływających na rokowanie odległe, co nie zawsze jest łatwe i wymaga sporego doświadczenia oraz wiedzy. Ponadto w przypadku istnienia kilku zmiennych istotnych w przewidywaniu analizowanego zdarzenia, wynik prezentowany jest za pomocą dość skomplikowanego wzoru, który czasami – zależnie od rodzaju analizowanych zmiennych – może być niezbyt przydatny w praktyce klinicznej.
Modele regresji dostarczają informacji o zmiennych prognostycznych dających jednakowy efekt rokowania odległego w całej badanej populacji. Lista uzyskanych ilorazów szans (OR) lub współczynników ryzyka (HR) jest zwykle pomocna w rankingu zmiennych w odniesieniu do ich istotności statystycznej, jednak wyjaśnienie ryzyka dla poszczególnych grup przypadków jest kłopotliwe. Nawet jeśli szacunkowe ryzyko wystąpienia niekorzystnego zdarzenia jest obliczane dla każdej kombinacji określonych parametrów, to wyniki nadal są tylko szacunkowe dla małych lub nawet hipotetycznych grup chorych [37]. Modele regresji koncentrują się na zmiennych o stosunkowo dużej istotności statystycznej, podczas gdy drzewa decyzyjne nie realizują globalnej optymalizacji dopasowania modelu do danych, ale w sposób sekwencyjny dzielą badaną populację na podgrupy, opierając się na najlepszej zmiennej prognostycznej. Pozwala to zidentyfikować w populacji podgrupy zagrożone wystąpieniem niekorzystnego zdarzenia oraz wydzielić chorych homogenicznych pod względem patofizjologii i przebiegu choroby [47]. W sposób automatyczny uzyskuje się ponadto zmienne istotne dla rokowania odległego w określonej podgrupie pacjentów, które w innej podgrupie mogą mieć znaczenie marginalne lub nawet nie wpływać na zaistnienie analizowanego zdarzenia. W efekcie zwiększa się dokładność i jakość prognostyczna modelu, który może być zastosowany w analizie konkretnego przypadku klinicznego. Pozwala to klinicyście zoptymalizować proces diagnostyczny i terapeutyczny oraz wskazuje zmienne kliniczne wymagające szczególnej uwagi w ocenie ryzyka powstania niekorzystnego zdarzenia [33]. Połączenie efektów rekursywnego podziału populacji z graficzną prezentacją w postaci drzewa (grafu), czyni z drzew decyzyjnych niezwykle przydatne narzędzie w modelowaniu ryzyka wystąpienia niekorzystnych zdarzeń w praktyce klinicznej [33,47]. Doniesienia literaturowe potwierdzają ponadto, że wyniki uzyskane w analizie z wykorzystaniem drzew decyzyjnych są bardziej szczegółowe i pozwalają lepiej wyjaśnić współzależności między analizowanymi wskaźnikami klinicznymi [49].

3.1. Współliniowość zmiennych oraz zmienne wikłające (*co-founders variables*)

Stosowanie modeli regresji wiąże się często z wielowspółliniowością (*multicolinearity*) [50]. W sytuacji, gdy zmienne wyjaśniające są wysoko skorelowane, wyniki analizy regresji mogą być niestabilne. Rozwiązaniem tego problemu może być przeformułowanie testowanego modelu regresji przez wyrażenie zmiennych prognostycznych będących w zależności liniowej, jako kompozyt tych zmiennych lub redukcję oryginalnego zbioru zmiennych prognostycznych do mniejszego i mniej skorelowanego podzbioru zmiennych. Wiąże się to jednak ze zreformułowaniem *a priori* postawionej hipotezy o związku między zmiennymi wyjaśniającymi a zmienną niezależną (wystąpieniem zdarzenia). Zagadnienie to jest kłopotliwe w realizacji (np. przez stosowanie metod krokowych, eliminacji wstecznej), ponadto uzyskane wyniki zależą od zastosowanych metod oraz doświadczenia badacza i zwykle nie są optymalne ani z teoretycznego, ani statystycznego punktu widzenia [50].
Niemniej ważnym aspektem, praktycznie pomijanym w dyskusjach dotyczących modeli regresji, jest tzw. paradoks Simpsona (inaczej efekt Yule-Simpsona), polegający na tym, że efekt działania kilku grup wydaje się odwrócony, kiedy są one połączone. Zjawisko to pojawia się wtedy, gdy zmienna ważona, która różni się od wartości określonej indywidualnie dla poszczególnych grup, jest używana do oceny połączonych grup [51,52,53,54,55]. Ma to miejsce w przypadku występowania tzw. zmiennych zakłócających (*co-founders*) lub w przypadku nierównomiernego rozłożenia badanej cechy między analizowanymi grupami.

3.2. Dane niepełne oraz koszty klasyfikacji

Niezwykle istotnym problemem związanym ze stosowaniem przedstawionych wcześniej klasycznych metod statystycznych jest konieczność analizy danych niepełnych (tzw. brakujących danych), które bardzo często pojawiają się w medycznych bazach danych. Rozróżnia się m.in. brakujące dane rozłożone losowo wśród wszystkich obserwacji (*missing completly at random* – MCAR) oraz brakujące dane zależne

od innych zmiennych (*missing at random* – MAR) [56]. Typowym rozwiązaniem omawianego zagadnienia jest usunięcie ze zbioru danych takiego przypadku (*listwise/pairwise deletion*) lub zastąpienie brakującej danej określoną wartością (średnia/mediana ze zbioru danych lub wartość uzyskana z założonego przez badacza modelu regresji wielowymiarowej; *imputation*) [57]. Metody imputacji nie są jednak popularne, ponieważ mogą prowadzić do uzyskania fałszywych wyników, np. w sytuacji, gdy zbyt duża liczba przypadków zostałaby usunięta. Powoduje to bowiem sztuczne zawyżenie błędu standardowego i zmniejszenie poziomu istotności testów (dla danych typu MCAR) oraz stronniczą estymację ze względu na niereprezentatywność badanej grupy (dla danych typ MAR) [58]. Uzupełnianie wartością średnią lub medianą w przypadku brakujących danych innych niż MCAR powoduje natomiast redukcję wariancji zmiennych i wtórnie wpływa na mniejsze wartości korelacji między zmiennymi. Wybór zmiennych do modelu regresji jest sprawą intuicyjną, a uzyskany wynik w dużej mierze zależy od doświadczenia i wiedzy badacza [56,57,58,59]. Należy jednak podkreślić, że w ciągu ostatnich kilku lat nastąpiło znaczące udoskonalenie metod imputacji, w związku z czym ich zastosowanie może stanowić alternatywę dla modeli drzew decyzyjnych.

Drzewa decyzyjne umożliwiają analizę zbioru danych w przypadku występowania danych niepełnych. W trakcie procesu uczenia drzewa z wykorzystaniem próby uczącej podział w danym węźle odbywa się wyłącznie z wykorzystaniem pełnych danych. Zgodnie z przyjętym kryterium, przeszukuje się przestrzeń zmiennych prognostycznych, aby znaleźć dla danej zmiennej wartość, dla której realizowany jest najlepszy podział podpróby do dwóch kolejnych węzłów lub liści. Po skonstruowaniu w danym węźle takiego podziału (zwanego właściwym), znajduje się w tym węźle kolejny podział, tzw. zastępczy (*surrogate*), który w podziale wykorzystuje inną zmienną, dającą podział podpróby w węźle możliwie najbardziej podobny do podziału właściwego. Możliwe jest zastosowanie kilku kolejnych zmiennych zastępczych. Pozwala to uwzględnić w procesie tworzenia drzewa zmienne, które w klasycznych metodach zostałyby usunięte z analizy. Oczywiście, zależnie od przyjętej miary różnorodności klas w węźle oraz algorytmu przeszukującego przestrzeń zmiennych, można w modelach drzew decyzyjnych uzyskać różne liczebności w kolejnych węzłach, mimo uzyskania tej samej zmiennej właściwej i tego samego punktu odcięcia w różnych drzewach [60].

W procesie tworzenia drzew decyzyjnych typu CART oraz LOTUS do optymalnego podziału wykorzystuje się tylko obserwacje z kompletnymi danymi. W procesie klasyfikacji w modelu CART, CRUISE i drzewie wnioskowania warunkowego wykorzystywane są natomiast tzw. podziały zastępcze (*surrogates splits*). Jeżeli zachodzi konieczność podziału w węźle na podstawie zmiennej *X*, o nieznanej wartości, stosuje się zmienną zastępczą *X'*, której wartością dysponujemy. Zmienną tę wybiera się w taki sposób, aby podział dokonany za jej pomocą, był możliwie zbliżony do podziału dokonanego za pomocą zmiennej [40,41,42,43]. W modelach QUEST oraz LOTUS brakujące wartości zmiennej są zastępowane średnimi obliczonymi z niebrakujących wartości danej zmiennej w klasie. W przypadku podziałów jednowymiarowych zastępowanie takie następuje dopiero w węźle, w którym brak wartości zmiennej wykorzystywanej do jego podziału. Natomiast dla podziałów wielowymiarowych zastępuje się brakujące wartości zmiennych już w korzeniu drzewa klasyfikacyjnego. W metodzie LOTUS możliwe jest również zastąpienie brakującej danej porządkowej modą danej klasy.

Niektóre algorytmy tworzące drzewo decyzyjne umożliwiają uwzględnienie w procesie uczenia niejednakowych kosztów błędnej klasyfikacji. W przypadku analizy klinicznej istotnym aspektem jest możliwość uwzględnienia faktu, że błędne zaklasyfikowanie chorego zagrożonego np. zgonem do grupy chorych niezagrożonych skutkuje dużo gorszymi implikacjami klinicznymi i społeczno-ekonomicznymi niż sytuacja odwrotna. Większość algorytmów umożliwia ponadto analizę danych nawet wtedy, gdy występują tzw. dane niepełne.

4. Złożoność drzewa klasyfikacyjnego

W przypadku modeli w postaci drzew decyzyjnych istotnym problemem jest wybór takiej postaci modelu, aby jego błąd predykcji był jak najmniejszy. Wraz ze złożonością modelu błąd resubstytucji maleje do zera, natomiast błąd dla zbioru testowego rośnie. Wystąpienie przeuczenia ogranicza zdolności generalizacyjne utworzonego drzewa, co może zaowocować błędnymi klasyfikacjami nowych przypadków. Ponadto zjawisko to powoduje znaczny rozrost struktury drzewa, które może sięgać np. kilkunastu poziomów [60]. Zbyt duże drzewo może prowadzić do nadmiernego dopasowania, a zbyt małe może mieć niedostateczną siłę przewidywania dla dokładnej klasyfikacji. Drzewa decyzyjne realizują to zabezpieczenie, opierając się na przycinaniu (*pruning*) polegającym na zastąpieniu wybranych poddrzew przez liście, którym przypisuje się etykietę kategorii najczęstszej wśród związanych z nim przykładów. Powoduje to redukcję wielkości drzewa przez usunięcie niektórych jego fragmentów oraz pogorszenie jakości klasyfikacji na zbiorze uczącym, ale daje zdecydowanie lepsze efekty w klasyfikacji nowych przypadków. Chroni to również drzewo przed nadmiernym „rozdrobnie-

niem". Efektem jest zmniejszenie stopnia złożoności modelu, czasami jednak kosztem usunięcia z niego niektórych zmiennych [60].
Zastosowanie klasycznych metod statystycznych w celu stworzenia wiarygodnych klinicznie reguł decyzyjnych bywa kłopotliwe dla badacza z następujących przyczyn:

- istnienie zwykle wielu możliwych zmiennych predykcyjnych utrudnia wybór tych właściwych,
- wiele zmiennych nie posiada rozkładu normalnego będącego podstawą parametrycznych metod statystycznych,
- w bazach biomedycznych zdarzają się braki danych, co utrudnia lub uniemożliwia zastosowanie tych przypadków w procesie tworzenia modelu decyzyjnego,
- w analizowanych bazach danych mogą istnieć złożone interakcje między zmiennymi oraz wzorce trudne do przeanalizowania; czasem – przy dużej liczbie zależności i zmiennych – modelowanie takich zależności nie jest w ogóle możliwe,
- stosowanie w praktyce wyników uzyskanych klasycznymi metodami bywa uciążliwe i nieintuicyjne dla lekarza klinicysty.

5. Przykłady zastosowań drzew klasyfikacyjnych w medycynie

Drzewa decyzyjne są powszechnie stosowane w medycynie i epidemiologii [46,47,48,61,62]. Takashi i wsp. [63] zastosowali drzewa typu CART w predykcji samoistnego krwawienia śródmózgowego. Jakość klasyfikacji drzewa decyzyjnego z 4 liśćmi, uzyskanego w grupie 347 chorych wyniosła 86%. Z kolei Long [64] porównał drzewa decyzyjne typu CART z regresją logistyczną w diagnostyce ostrego zespołu wieńcowego u 3453 chorych, wykazując przydatność obu metod i uzyskując dokładność klasyfikacji 76% (przy czułości 63% i specyficzności 81%). Zastosowanie drzew decyzyjnych typu CART w przewidywaniu wystąpienia zgonu u 9484 chorych po zawale przedstawił Austin [65], a Negassa i wsp. ocenili ryzyko zgonu wewnątrzszpitalnego po wykonaniu przezskórnej angioplastyki naczyń wieńcowych u 5385 chorych z ostrym zespołem wieńcowym (uzyskana jakość wyniosła 82%) [65]. Misztal przedstawił zastosowanie metod rozpoznawania obrazów (w tym drzew decyzyjnych) we wspomaganiu diagnostyki medycznej [66], zaś Trzpiot i wsp. zaprezentowali zastosowanie drzew decyzyjnych w naukach społecznych, obejmujące ocenę zmian zachowań komunikacyjnych uwarunkowanych znajomością alternatywnych metod pokonywania przestrzeni miejskiej [67,68].

6. Podsumowanie

Zastosowanie w praktyce klinicznej procesu wspomagania oceny ryzyka opartego na drzewie decyzyjnym może ułatwić podejmowanie decyzji co do sposobu postępowania terapeutycznego w wybranych jednostkach chorobowych oraz przynieść wymierne korzyści społeczno-ekonomiczne.

Praca naukowa sfinansowana w ramach badań statutowych w latach 2012–2013, nr umowy KNW-1-008/P/2/0.

PIŚMIENNICTWO

1. Garg A.X., Adhikari N.K.J., McDonald H. et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. A Systematic Review. JAMA 2005; 293: 1223–1238.
2. Lindgaard G., Pyper C., Frize M., Walker R. Does Bayes have it? Decision Support Systems in diagnostic medicine. Int. J. Ind. Ergon. 2009; 39: 524–532.
3. Tierney W. Improving clinical decision and outcomes with information: a review. Int. J. Med. Inf. 2001; 62: 1–9.
4. Berlin A., Sorani M., Sim I. A taxonomic description of computer-based clinical decision support systems. J. Biomed. Info. 2006; 39: 656–667.
5. Montgomery A.A., Fahey T., Peters T.J., MacIntosh C., Sharp D.J. Evaluation of computer based clinical decision support system and risk chart for management of hypertension in primary care: randomised controlled trial. BMJ 2000; 320: 686–690.
6. Kawamoto K., Houlihan C.A., Balas E.A., Lobach D.F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005; 330: 765–772.
7. Reisman Y. Computer-based clinical decision aids. A review of methods and assessment of systems. Med. Inform. 1996: 21: 179–197.
8. Zupan B., Porenta A., Vidmar G., Aokin N., Bratko I., Beck J.R. Decision at hand: a decision support system on handhelds. Stud. Health Technol. Inform. 2001; 84: 566–570.
9. Bagley S.C., White H., Golomb B.A. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J. Clin. Epidemiol. 2001; 54: 979–985.
10. Long W.J., Griffith L.J., Selker H.P., D'Agostino R.B. A comparison of logistic regression to decision-tree induction in a medical domain. Comput. Biomed. Res. 1993; 26: 74–97.
11. Huang D., Quan Y., He M., Zhou B. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. J. Exp. Clin. Cancer Res. 2009; 28: 149–156.
12. Polat K., Günes S. A hybrid medical decision making system based on principles component analysis, k-NN based weighted pre-processing and adaptive neuro-fuzzy inference system. Dig. Sig. Proc. 2006; 16: 913–921.
13. Ji S.Y., Smith R., Huynh T., Najarian K. A comparative analysis of multi-level computer-assisted decision making systems for traumatic injuries. BMC Med. Inform. Decis. Mak. 2009; 9: 2–18.
14. Arif M., Akram M.U., Minhas F.A. Pruned fuzzy K-nearest neighbor classifier for beat classification. J. Biomed. Sci. Eng. 2010; 3: 380–389.
15. Verplancke T., Van Looy S., Benoit D. et al Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. BMC Med. Inform. Decis. Mak. 2008; 8: 56–63.
16. Chen S.T., Hsiao Y.H., Huang Y.L. et al. Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power Doppler imaging. Korean J. Radiol. 2009; 10: 464–471.
17. Lisboa P.J., Taktak A.F.G. The use of artificial neural networks in decision support in cancer: A systematic review. Neural. Netw. 2006; 19: 408–415.
18. Glaser J. Clinical decision support: the power behind the electronic health record. Healthc. Financ. Manage 2008; 62(7): 50–51.
19. Lenz R., Reuchert M. IT support for healthcare process – premises, challenges, perspectives. Data Knowl. Eng. 2007; 61: 39–58.

20. Bairstow P., Persaud J., Mendelson R., Ngyuen L. Reducing inappropriate diagnostic practice through education and decision support. Int. J. Qual in Health Care 2010; 22(3): 194–200.
21. Haynes R.B., Wilczyński N.L. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: Methods of a decision-maker-research partnership systemic review. Implement. Sci. 2010; 5: 5–12.
22. Thursky K.A., Buising K.L., Bak N. et al. Reduction of broad-spectrum antibiotic use with computerized decision support in an intensive care unit. Int. J. Qual Health Care 2006; 18(3): 224–231.
23. Delpierre C., Cuzin L., Fillaux J., Alvarez M., Massip P., Lang T. A systematic review of computer-based patient record systems and quality of care: more randomized clinical trials or a broader approach? Int. J. Qual Health Care 2004; 16(5): 407–416.
24. Sintchenko V., Iredell J.R., Gilbert G.L., Coiera E. Handheld computer-based decision support reduces patient length of stay and antibiotic prescribing in critical care. J. Am. Med. Inform. Assoc. 2005; 12(4): 398–402.
25. Durieux P., Nizard R., Ravaud P., Mounier N., Lepage E. A clinical decision support system for prevention of venous thromboembolism: effect on physician behavior. JAMA 2000; 283: 2816–2821.
26. Leslie S.J., Denvir M.A. Clinical decision support software for chronic heart failure. Crit. Pathw. Cardiol. 2007; 6: 121–126.
27. Dolan J.G. Shared decision-making – transferring research into practice: the Analytic Hierarchy Process (AHP). Patient. Educ. Couns. 2008; 73: 418–425.
28. Hardy D., Smith B. Decision making in clinical practice. Brit. J. Anaest. Rec. Nurs. 2008; 9: 19–21.
29. Wennberg J.E. Improving the medical decision-making process. Health. Aff. 1988; 7: 99–106.
30. Bates D.W., Kuperman G.J., Wang S. et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J. Am. Med. Inform. Assoc. 2006; 10: 523–530.
31. Lin C., Lin CM., Lin B., Yang M.C. A decision support system for improving doctor's prescribing behavior. Expert. Syst. Appl. 2009; 36: 7975–7984.
32. Young S.A., Chaney E., Shoai R. et al. Information technology to support improved care for chronic illness. J. Gen. Inter. Med. 2007; 22: 425–430.
33. Sutton C.D. Classification and regression trees, bagging and boosting. In: Handbook of Statistics. Ed. Pfefferman D. Elsevier New York. 2010, 303–327.
34. Breiman L., Friedman J., Stone C.J., Olshen R.A. Classification and regression trees. Chapman and Hall/CR, New York 1993.
35. Quinlan J.R. C4.5: Programs for machine learning. Morgan Kauffman, London 1993.
36. Koronacki J., Ćwik J. Statystyczne systemy uczące się. Oficyna Wydawnicza Exit, Warszawa 2008.
37. Rokach L., Maimon O. Data mining with decision trees. World Scientific Publishing Co. Pte. Ltd. Singapore 2008.
38. Almuallim H., Kaneda S., Akiba Y. Development and applications of decision trees. Expert Systems 2002; 1: 53–77.
39. Kotsiantis S.B. Supervised machine learning: a review of classification techniques. Informatica 2007; 31: 249–268.
40. Krzyśko M., Wołyński W., Górecki T., Skorzybut M. Systemy uczące się. Wydawnictwo WNT, Warszawa 2009.
41. Podgorelec V., Kokol P., Stiglic B., Rozman I. Decision trees: an overview and their use in medicine. J. Med. Sys. 2002; 26: 445–463.
42. Loh W.Y., Shih Y.S. Split selection methods for classification trees. Stat. Sin. 1997; 7: 815–840.
43. Esposito F., Malerba D., Semeraro G. A comparative analysis of methods for pruning decision trees. Machine Learning 1997; 19: 476–491.
44. Hothorn T., Hornik K., Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. J. Comp. Graphical. Stat. 2006; 15: 651–674.
45. Lucas P.J.F., Abu-Hanna A. Prognostic methods in medicine. Artif. Intell. Med. 1999; 15: 105–119.
46. Camdeviren H.A., Yazici A.C., Akkus Z., Bugdayci R., Sungur M.A. Comparison of logistic regression model and classification tree: an application to postpartum depression data. Expert Sys. Appl. 2007; 32: 987––994.
47. Mello F.C., Valle Baastos L.G., Soares S.L.M. et al. Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study. BMC Public Health 2006; 6(43): 1–8.
48. Zhang H., Holford T., Bracnek M.B. A tree-based method of analysis for prospective studies. Stat. Med. 1996; 15: 37–49.
49. Banerjee M., George J., Song EY., Roy A., Hryniuk W. Tree-based model for breast cancer prognostication. J. Clin. Oncol, 2004; 22: 2567–2574.
50. Freckleton R.P. Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav. Ecol. Sociobiol. 2011; 65: 91–101.
51. Marshall R.J. The use of classification and regression trees in clinical epidemiology. J. Clin. Epid. 2001; 54: 603–609.
52. Long W. A comparison of logistic regression to decision-tree induction in a medical domain. Comp Biomed. Res. 1993; 26: 74–97.
53. Tu Y.U., Gunnel D., Gilthorpe M.S. Simpson's paradox, lord's paradox, and suppression effect are the same phenomenon – the reversal paradox. Emerg. Themes Epidemiol. 2008; 5(2): 1–9.
54. Rücker G., Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. BMC Med. Res. Methodol. 2008; 8: 1–8.
55. Ameringer S., Serlin RC., Ward S. Simpson's paradox and experimental research. Nurs. Res. 2009; 58: 123–127.
56. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. John Wiley and Sons, New York 2002.
57. Scheffer J. Dealing with missing data. Res. Lett. Inf. Math. Scil. 2002; 3: 153–160.
58. Schafer J.L., Graham J.W. Missing data: our view of the state of the art. Psychol. Methods 2002; 7: 147–177.
59. Muller R., Möckel M. Logistic regression and CART in the analysis of multimarker studies. Clin. Chim. Acta 2008; 394: 1–6.
60. Walesiak M., Gatnar E. Statystyczna analiza danych z wykorzystaniem programu R. Wydawnictwo Naukowe PWN, Warszawa 2009.
61. Zhang H., Holford T., Bracnek M.B. A tree-based method of analysis for prospective studies. Stat. Med. 1996; 15: 37–49.
62. Mello F.C., Valle Baastos L.G., Soares S.L.M. Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study. BMC Public Health 2006; 6(43): 1–8.
63. Takashi O., Cook EF., Nakamura T., Saito J., Ikawa F., Fukui T. Risk stratification for in-hospital mortality in spontaneous intracerrebral haemorrhage: A classification and regression tree analysis. QJM 2006; 99: 743–750.
64. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat. Med. 2007; 26: 293–2957.
65. Negassa A., Monrad E.S., Bang J.Y., Vankeepuram S.S. Tree-structured risk stratification of in-hospital mortality after percutaneous coronary intervention for acute myocardial infarction: a report from the New York State percutaneous coronary intervention database. Am. Heart J. 2007; 154: 321––329.
66. Misztal M. O zastosowaniu statystycznych metod rozpoznawania obrazów do wspomagania procesów podejmowania decyzji w diagnostyce medycznej. StatSoft Polska 2003. http://www.statsoft.pl/czytelnia/ badania-naukowe/d1biolmed/ obrazy.pdf
67. Szołtysek J., Trzpiot G. Drzewa klasyfikacyjne w badaniu preferencji komunikacyjnych. W: Studia Ekonomiczne nr 97: Modelowanie preferencji a ryzyko '12. Red. T. Trzaskalik. Zesz. Nauk. Wydziałowe Uniwersytetu Ekonomicznego w Katowicach, Katowice 2012; 213–230.
68. Trzpiot G., Ganczarek A. Drzewa decyzyjne w statystycznej analizie decyzji na przykładzie wirtualnych łańcuchów dostaw. FOE 2012; 271: 57–70.