Przeprowadzanie segmentacji przedsiębiorstw za pomocą drzew klasyfikacyjnych

Mirosława Lasek, profesor
Katedra Informatyki Gospodarczej i Analiz Ekonomicznych, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski
Marek Pęczkowski, mgr
Katedra Informatyki Gospodarczej i Analiz Ekonomicznych, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski

Wstęp

Drzewa klasyfikacyjne (zwane także drzewami decyzyjnymi) są jedną z metod eksploracji danych (ang. *Data Mining*). W metodach eksploracji danych nie zakłada się znajomości rozkładów statystycznych cech ani postaci analitycznej związku między cechami, jak np. w analizie regresji. Nie trzeba więc weryfikować wcześniej przyjętych założeń dotyczących rozkładów zmiennych.

W tej pracy podjęto próbę wykorzystania drzew klasyfikacyjnych do segmentacji przedsiębiorstw — polskich spółek giełdowych — na podstawie danych określających ich sytuację finansową i majątkową. Jest to kontynuacja prac prowadzonych w ramach realizowanych od kilku lat badań statutowych w Katedrze Informatyki Gospodarczej i Analiz Ekonomicznych WNE UW, i dotyczących zastosowania zaawansowanych metod eksploracji danych do analizy i oceny kondycji polskich firm. Zmienne używane do klasyfikowania przedsiębiorstw zostały zamieszczone w załączniku. Są to wskaźniki finansowe i majątkowe, wskaźniki pozycji na rynku kapitałowym oraz wielkości ze sprawozdań finansowych (bilans, rachunek zysków i strat, rachunek przepływu środków pieniężnych). Dane te zaczerpnięto z opracowań *Wyniki finansowe spółek giełdowych*, wydawanych kwartalnie przez firmę Notoria Serwis.

Przedmiotem metod eksploracji danych jest wykrywanie związków i relacji występujących w dużych zbiorach danych [Kovalerchuk, Vityaev, 2000; Shi, 2000; Witten, Frank, 2000; Zakrzewicz, 2001]. Stosuje się takie metody jak analiza skupień, sieci neuronowe, algorytmy genetyczne, generowanie reguł rozmytych, różne metody graficznej wizualizacji danych (np. analizę korespondencji), a także opisywane w tym artykule drzewa klasyfikacyjne. Metody te rozwinęły się wraz z rozwojem techniki komputerowej, ponieważ do ich realizacji potrzeba dużej mocy obliczeniowej — ze względu na złożoność algorytmów i duży czas obliczeń.

Drzewa klasyfikacyjne stanowią graficzną reprezentację metody rekurencyjnego podziału [Gatnar, 2001]. Metoda rekurencyjnego podziału polega na

stopniowym (hierarchicznym) podziale wielowymiarowej przestrzeni cech (w której znajduje się zbiór obiektów) na rozłączne podzbiory aż do osiągnięcia ich jednorodności ze względu na wyróżnioną cechę — nazywaną zmienną objaśnianą. W praktyce proces podziału jest często zatrzymywany wcześniej, aby uniknąć tworzenia podzbiorów o bardzo małej liczbie elementów.

Przedmiotem klasyfikacji jest zbiór obiektów S, charakteryzowanych przez $(m + 1)$ — wymiarowy wektor cech $(x_1, ..., x_m, y)$. Zmienna y jest wyróżnioną zmienną, ze względu na którą dokonujemy klasyfikacji (zmienna objaśniana).

W zależności od rozwiązywanego problemu y może być zmienną jakościową — nominalną lub porządkową, albo zmienną ciągłą — przyjmującą wartości ze zbioru liczb rzeczywistych. W szczególności y może być zmienną binarną (przyjmującą tylko dwie różne wartości np. 0 i 1).

Wśród zmiennych $x_1, ..., x_m$ (objaśniających) również mogą występować zmienne jakościowe oraz ilościowe.

Dysponując n obserwacjami wszystkich zmiennych $y, x_1, ..., x_m$ szukamy relacji między zmiennymi dającej opisać się przez model postaci $y = f(x_1, ..., x_m,; \beta_1, ... \beta_k) + \varepsilon$, gdzie β_j $(j = 1, ..., k)$ są parametrami modelu, a ε — składnikiem losowym.

Jeżeli cecha y jest jakościowa, to mamy do czynienia z modelem dyskryminacyjnym (klasyfikacyjnym). Jeżeli cecha y jest ilościowa, to mamy do czynienia z modelem regresyjnym.

Celem analizy dyskryminacyjnej jest znalezienie charakterystyki podzbiorów, na które podzieliliśmy zbiór obiektów S. Podzbiory te będziemy nazywać też klasami. Otrzymane wyniki powinny pozwolić na przewidywanie przynależności do klas nowych obiektów (spoza zbioru S). Zbiór, na którego podstawie tworzymy klasy, nazywa się zbiorem uczącym. Uzyskane wyniki podziału na klasy na podstawie zbioru uczącego stosujemy zatem do klasyfikacji obiektów, których przynależność do klas nie jest znana.

W metodach regresyjnych celem jest znalezienie związku opisującego wpływ jednej lub wybranej liczby cech spośród $x_1, ..., x_m$ na cechę y.

1. Drzewa klasyfikacyjne jako graficzna ilustracja rekurencyjnego podziału zbioru obiektów

Jak zaznaczyliśmy we wstępie, proces rekurencyjnego podziału zbioru obiektów na rozłączne podzbiory można graficznie przedstawić w postaci drzewa. Budowa drzewa odzwierciedla sekwencję kolejnych kroków podziału.

Drzewa klasyfikacyjne ilustrują podział zbioru obiektów na kolejne, hierarchicznie uporządkowane podzbiory (klasy), który dokonywany jest aż do osiągnięcia warunków określających możliwie jednorodną przynależność obiektów do klas lub gdy spełniony zostaje ustalony warunek zakończenia klasyfikacji. Warunkiem tym może być na przykład maksymalna wartość określająca liczbę poziomów drzewa (osiągnięcie maksymalnej „głębokości

drzewa") lub osiągnięcie minimalnej liczebności węzłów podlegających podziałowi.

Posługując się terminami z teorii grafów, można powiedzieć, że drzewa klasyfikacyjne są grafami spójnymi, niezawierającymi cykli, a więc takimi, w których istnieje tylko jedna droga między dwoma różnymi wierzchołkami, nazywanymi też węzłami.

Mówimy, że graf jest spójny, jeżeli dla dowolnej pary wierzchołków w_i, w_j $(i \neq j)$ istnieje droga z w_i do w_j, tzn. skończony ciąg krawędzi (w_1, w_2), (w_2, w_3), ... (w_{l-1}, w_l), gdzie $w_1 = w_i$ oraz $w_l = w_j$. Jeżeli $w_1 = w_l$, to drogę nazywamy cyklem. Jeżeli graf spójny nie zawiera cykli, to można go przedstawić w postaci struktury hierarchicznej, zwanej drzewem. Wierzchołek początkowy drzewa, z którego wychodzą przynajmniej dwie krawędzie, nazywamy korzeniem drzewa. Wierzchołek końcowy, z którego nie wychodzą żadne krawędzie, nazywamy liściem drzewa.

Najdłuższą drogę — ze względu na liczbę krawędzi tworzących tę drogę — między korzeniem a dowolnym liściem nazywamy głębokością drzewa. Liczba liści określa wielkość drzewa.

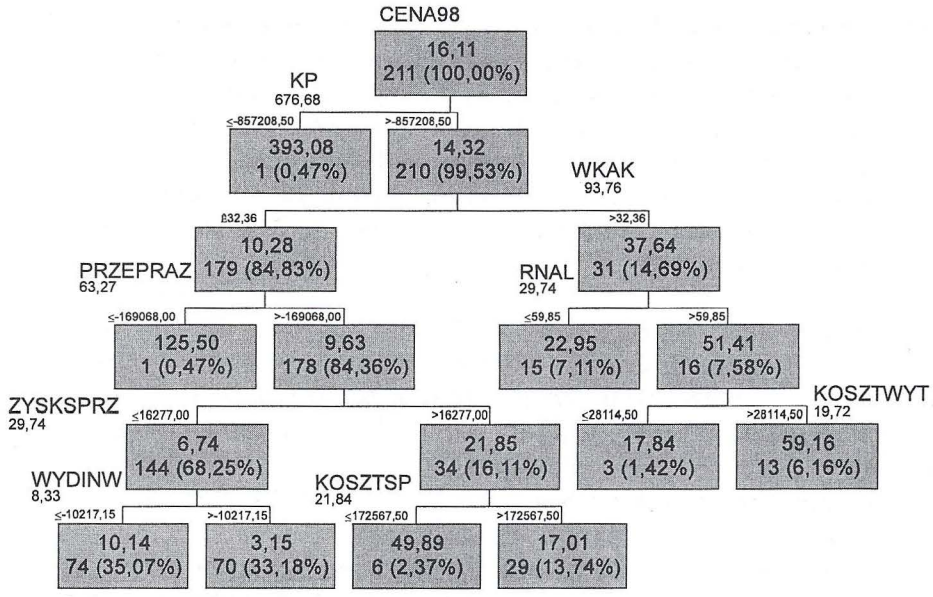
Drzewo, w którym z każdego wierzchołka wychodzą dwie krawędzie, nazywamy drzewem binarnym.

Tak konstruowane drzewa klasyfikacyjne umożliwiają przedstawianie procesu podziału zbioru przedsiębiorstw na jednorodne klasy, charakteryzowane określonymi wartościami atrybutów. Wewnętrzne wierzchołki określają sposób dokonywania podziału — na podstawie wartości cech obiektów — w naszym przypadku przedsiębiorstw. Liście reprezentują klasy, do których należą obiekty-przedsiębiorstwa. Krawędzie drzewa wskazują wartości cech, na których podstawie dokonywany jest podział. Przykład prostego drzewa klasyfikacji przedsiębiorstw przedstawiono na rysunku 1.

Na rys. 1. zawarte są informacje pozwalające odczytać reguły przynależności przedsiębiorstw do klas. Na przykład dla drzewa przedstawionego na rysunku 1., wybierając za każdym razem gałąź idącą na prawo, utworzymy regułę:

Jeżeli KP > -857208,50 i WKAK > 32,36 i RNAL > 59,85 i KOSZTWYT > 28114,50, to CENA98 wynosi średnio 59,16 (tak jest w przypadku 13 przedsiębiorstw, tj. 6% wszystkich analizowanych przedsiębiorstw).

Możemy zauważyć, że w zbiorze istnieje jedno przedsiębiorstwo o dużym ujemnym kapitale pracującym, ale wysokiej cenie akcji (393 zł). Przedsiębiorstwa o wyższej wartości księgowej na jedną akcję mają zdecydowanie wyższy średni kurs akcji. Jest ich jednak znacznie mniej niż przedsiębiorstw o niższej wartości księgowej na jedną akcję.



Rys. 1.

Przykład drzewa klasyfikacji przedsiębiorstw o różnym średnim kursie akcji uzyskany na podstawie analizy wskaźników finansowych i majątkowych

Oznaczenia:
CENA98 — średni kurs akcji w roku 1998,
KP — kapitał pracujący,
PRZEPRAZ — przepływy pieniężne razem,
WKAK — wartość księgowa na jedną akcję,
RNAL — rotacja należności w dniach,
ZYSKSPRZ — zysk (strata) na sprzedaży,
KOSZTWYT — koszty wytworzenia sprzedanych produktów,
WYDINW — wydatki z tytułu działalności inwestycyjnej,
KOSZTSP — koszty sprzedanych produktów, towarów i materiałów.
Źródło: na podstawie wyników z programu AnswerTree.

2. Rozwój algorytmów tworzenia drzew klasyfikacyjnych

Początki zainteresowania się metodami rekurencyjnego podziału sięgają lat 60. XX w. Jednym z pierwszych algorytmów tworzenia drzew klasyfikacyjnych był powstały w tym czasie CLS (ang. *Concept Learning System*). Dużą rolę w rozwoju algorytmów tworzenia drzew klasyfikacyjnych odegrał algorytm ID3 opracowany przez Quinlana (1983) [Gatnar, 1998]. Algorytm ten był intensywnie rozwijany i modyfikowany. W 1996 r. pojawiła się jego znana wersja C4.8 [Quinlan, 1996]. Algorytm Quinlana był intensywnie wykorzystywany na Wydziale Nauk Ekonomicznych UW, gdzie opracowano m.in. program komputerowy umożliwiający generowanie reguł, pozwalających wnioskować o kondycji przedsiębiorstw na podstawie analiz sprawozdań finansowych. Wyniki prac zostały opisane m.in. w [Lasek, Pęczkowski, 1991].

2.1. Ogólny schemat postępowania dla utworzenia drzewa klasyfikacyjnego

Zastosowane procedury polegają na sukcesywnym dzieleniu zbioru obiektów S na podzbiory, aż do osiągnięcia maksymalnej ich jednorodności pod względem wartości zmiennej zależnej (objaśnianej). W zależności od rodzaju zmiennej zależnej (cecha nominalna, porządkowa, ciągła) stosuje się różne miary do oceny jednorodności. Podział zbioru obiektów na podzbiory odbywa się w kolejnych krokach na podstawie wartości wybranych zmiennych niezależnych (objaśnianych).

Ogólny schemat postępowania dla utworzenia drzewa klasyfikacyjnego przedstawiono m.in. w pracy [Gatnar, 2001]. Tworzenie drzewa klasyfikacyjnego odbywa się na podstawie zbioru uczącego, w którym przynależność obiektów do poszczególnych klas jest znana. Zgodnie z opisem, przedstawionym w pracy [Gatnar, 1998; Gatnar, 2001] schemat budowy drzewa wygląda następująco:

1. Mając dany zbiór obiektów S sprawdź, czy należą do tej samej klasy. Jeżeli tak, to zakończ postępowanie.
2. W przeciwnym przypadku rozważ wszystkie możliwe podziału zbioru S na rozłączne podzbiory $S_1, S_2, \ldots, S_s$ tak, by były jak najbardziej jednorodne (s — liczba podzbiorów).
3. Dokonaj oceny jakości każdego z tych podziałów zgodnie z przyjętym kryterium i wybierz najlepszy z nich.
4. Podziel zbiór S w wybrany sposób.
5. Wykonaj kroki 1–4 rekurencyjnie, przyjmując jako S każdy z otrzymanych podzbiorów $S_1, S_2, \ldots, S_s$.

Procedura podziału może zostać zakończona, jeżeli zostało osiągnięte jedno z przyjętych kryteriów zakończenia (tzw. kryterium stopu). Przedstawionym powyżej krokom procesu podziału rekurencyjnego odpowiadają kolejne kroki budowy hierarchicznego drzewa klasyfikacyjnego. Budowa drzewa odzwierciedla sekwencję kolejnych kroków procedury.

2.2. Algorytmy tworzenia drzew klasyfikacyjnych

Opracowano wiele algorytmów szczegółowych realizujących przedstawiony powyżej ogólny schemat budowy drzew klasyfikacyjnych [*AnswerTree. User's Guide*, 1998; http://www.spss.pl]. Algorytmy te różnią się sposobem wyboru cech, na których podstawie następuje podział zbioru obiektów, kryterium zakończenia podziału powstającego podzbioru obiektów, sposobem przydzielania obiektów znajdujących się w liściu drzewa do określonej klasy, postacią funkcji oceniającej jakość podziału, sposobem klasyfikacji obiektów o brakujących wartościach cech.

Podziału przestrzeni cech, w której znajdują się obiekty, nie dokonuje się w sposób przypadkowy, ale zgodnie z pewnym kryterium — jest nim pewna funkcja oceniająca jakość podziału (stopień jednorodności podzbiorów), która jest maksymalizowana. Zamiast szukać maksimum tej funkcji, można szukać minimum pewnej funkcji H mierzącej niejednorodność (tzn. heteroge-

niczność), zwanej miarą zanieczyszczenia (ang. *impurity*). Stosuje się różne miary zanieczyszczenia, w zależności od tego, czy cecha zależna jest jakościowa, czy ilościowa. Warunki, jakie powinna spełniać miara zanieczyszczenia, i przegląd miar podaje np. Gatnar [2001]. Jedną z powszechnie stosowanych miar jest miara χ^2, obliczana na podstawie tablicy kontyngencji postaci:

	S_1	S_1	.	.	.	S_j	.	.	.	S_s	
y_1	n_{11}	n_{12}	.	.	.					n_{1s}	$n_{1.}$
y_2	n_{21}										$n_{2.}$
.											
.											
.											
y_i						n_{ij}					
.											
.											
.											
y_r	n_{r1}									n_{rs}	$n_{r.}$
	$n_{.1}$					$n_{.j}$				$n_{.s}$	n

(1)

gdzie:

S_j — podzbiór zbioru obiektów S (j = 1, ..., s),

y_i — wartość cechy y (i = 1, ... r),

n_{ij} — liczba obiektów mających wartość cechy $y = y_i$ i należących do podzbioru S_j.

$n_{i.} = \sum_{j=1}^{s} n_{ij}$, $n_{.j} = \sum_{i=1}^{r} n_{ij}$ — są liczebnościami brzegowymi.

Zachodzi przy tym

$$n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \tag{2}$$

Oznaczając liczebności teoretyczne w komórkach tablicy kontyngencji:

$$\hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \tag{3}$$

miarę χ^2 definiujemy jako:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} \tag{4}$$

Wartość

$$p_i = \frac{n_{i.}}{n} \tag{5}$$

oznacza prawdopodobieństwo, że losowo wybrany obiekt ma wartość cechy $y = y_i$ (tzn., że należy do i-tej klasy ze względu na cechę y). Prawdopodobieństwo to jest nazywane prawdopodobieństwem *a priori*.

W algorytmach tworzenia drzew klasyfikacyjnych prawdopodobieństwa *a priori* odgrywają ważną rolę. Ich wartości określa się np. na początku procedury podziału w programach komputerowych realizujących algorytmy tworzenia drzew klasyfikacyjnych. Te prawdopodobieństwa mogą być:

- szacowane na podstawie zbioru danych — zbioru uczącego, według wzoru (5),
- przyjmowane jako równe, tzn. $p_1 = p_2 = \ldots = p_r$,
- szacowane na podstawie zbioru testowego,
- ustalane arbitralnie jako wartości wskazane przez użytkownika.

Prawdopodobieństwo, że w wyniku podziału pewien losowo wybrany obiekt znajdzie się w podzbiorze S_j wynosi:

$$q_j = \frac{n_{.j}}{n} \tag{6}$$

Takie prawdopodobieństwo wraz z przyjętą miarą heterogeniczności H wykorzystujemy do poszukiwania najlepszego podziału na danym etapie procedury.

Najczęściej na danym etapie (w danym wierzchołku drzewa) podział następuje na podstawie wartości tylko jednej zmiennej objaśniającej x_t ($1 \leq t \leq m$). Oznacza to podział przestrzeni cech hiperpłaszczyznami równoległymi do osi. Czasami obiekty są tak ułożone w przestrzeni wielowymiarowej, że podział przestrzeni hiperpłaszczyznami równoległymi do osi aż do uzyskania podzbiorów homogenicznych jest długotrwały i prowadzi do drzew o dużej głębokości. Wtedy lepsze efekty może dać podział hiperpłaszczyznami ukośnymi względem osi, tzn. podział na podstawie kombinacji liniowych kilku cech niezależnych.

Szczególnym przypadkiem — często stosowanym — jest podział na dwie części, gdy w ustalonym wierzchołku drzewa następuje podział zbioru na dwa podzbiory. Prowadzi to do utworzenia drzewa binarnego.

Wierzchołek do podziału jest wybierany w ten sposób, aby maksymalizować spadek poziomu heterogeniczności:

$$\Delta H(S, x_t) = H(S) - \sum_{j=1}^{s} H(S_j) q_j \tag{7}$$

W podanym powyżej wzorze $H(S)$ — oznacza stopień heterogeniczności zbioru przed danym podziałem ($S = S_1 \cup S_2 \cup \ldots S_s$), $H(S_j)$ — stopień heterogeniczności zbioru S_j, powstałego po podzieleniu zbioru S, q_j — jest prawdo-

podobieństwem, wyznaczonym zgodnie ze wzorem (6), $\Delta H(S, X_t)$ — zmiana heterogeniczności po dodaniu podziału na podstawie zmiennej x_t. Szczególnym przypadkiem jest podział na dwa podzbiory, gdy $s = 2$.

Na ogół proces wyboru zmiennej x_t i sposobu, w jaki nastąpi podział zbioru na podzbiory, odbywają się jednocześnie. Można jednak postąpić inaczej, tj. najpierw wybrać najlepszą zmienną, a następnie szukać optymalnego dla niej podziału zbioru wartości. Taka metoda została zastosowana w algorytmie QUEST [Loh, Shih, 1997].

Jeżeli x_t jest zmienną nominalną przyjmującą s różnych wartości, to podział danego zbioru S na podstawie tej zmiennej tworzy s podzbiorów (S_1, ..., S_s). Prowadzi to do utworzenia zbiorów, które są mało liczne. Ponadto miary heterogeniczności są na ogół wrażliwe na liczbę wariantów cech i preferują zwykle te zmienne, które mają więcej różnych wartości. Rozwiązaniem tego problemu jest łączenie wartości zmiennej w grupy. Najczęściej tworzy się dwie grupy. Jedną z możliwości jest utworzenie dwóch podzbiorów: w pierwszym znajduje się pojedyncza wartość zmiennej $x_t = x_{t_0}$, a w drugiej wszystkie pozostałe wartości tej zmiennej. Najczęściej jednak dokonuje się najpierw podziału na podzbiory odpowiadające pojedynczym wartościom zmiennej x_t, a później te podzbiory stopniowo się łączy. Dla zmiennych porządkowych należy zadbać o to, żeby łączone były tylko sąsiednie wartości. Jednocześnie sprawdza się statystyczną istotność każdego z wariantów, np. w metodzie CHAID wybiera się taki podział, który daje największą istotność sprawdzianu testu χ^2 liczonego dla tablicy kontyngencji [Biggs, de Ville, Suen 1991; Kass, 1980].

Zmienne ciągłe x_t muszą być poddane procesowi dyskretyzacji, tzn. podziałowi uporządkowanego zbioru wartości zmiennej na T przedziałów postaci $<v_j, v_{j+1})$, gdzie $v_0 < v_1 < v_2 < \ldots < v_T$ (T — jest pewną ustaloną liczbą całkowitą). W najprostszym przypadku zbiór wartości zmiennej x_t zostaje podzielony na dwa przedziały postaci $x_t \leq C$ i $x_t > C$, gdzie C jest pewną stałą. Stosuje się różne elementy szukania optymalnego punktu podziału C. Na ogół w algorytmach wykorzystuje się znajomość wartości zmiennej zależnej, tak aby podzbiory były jak najbardziej homogeniczne ze względu na zmienną zależną.

3. Programy komputerowe tworzenia drzew klasyfikacyjnych

Metody rekurencyjnego podziału zostały zaimplementowane w wielu pakietach i programach komputerowych, np.:

- STATISTICA — moduł „Drzewa klasyfikacyjne" (tu znajdują się trzy metody służące do budowania modeli dyskryminacyjnych).
- S-PLUS — jest to pakiet statystyczny firmy MathSoft, ma procedury służące do budowy drzew klasyfikacyjnych i regresyjnych. Niektóre procedury są dostępne w Internecie.
- CART — realizuje algorytm podany w pracy: [Breiman, Friedman, Olshen, Stone, 1984].

- MARS — firmy Salford Systems, wykorzystywany do budowy modeli regresyjnych, opartych na krzywych składanych.

Jednym z najbardziej uniwersalnych programów tworzenia drzew klasyfikacyjnych jest AnswerTree firmy SPSS [*AnswerTree. User's Guide*, 1998; http://www.spss.pl], współpracujący z pakietem SPSS for Windows. Może być też zakupiony i używany jako samodzielnie pracujący program. Zawiera on cztery procedury: CHAID, Exhaustive CHAID, C&RT, QUEST.

3.1. Algorytmy zastosowane do przeprowadzenia segmentacji przedsiębiorstw za pomocą drzew klasyfikacyjnych

Na potrzeby segmentacji przedsiębiorstw wykorzystywaliśmy program AnswerTree i udostępniane przez ten program cztery wymienione algorytmy tworzenia drzew klasyfikacyjnych:

1) CHAID (*Chi-squared Automatic Interaction Detection*) [Biggs, de Ville, Suen, 1991; Kass, 1980];
2) Exhaustive CHAID;
3) C&RT (*Classification & Regression Trees*) [Breiman et al., 1984];
4) QUEST (*Quick, Unbiased, Efficient, Statistical Tree*) [Loh, Shih, 1997].

Wymienione powyżej algorytmy umożliwiają rozważanie różnego typu zmiennych zależnych — objaśnianych (ang. *target variables*)[1], jak i różnego typu zmiennych niezależnych — objaśniających, zwanych także predyktorami (ang. *predictor variables*). Możemy rozważać zmienne nominalne, porządkowe i ciągłe.

Wszystkie wymienione powyżej algorytmy umożliwiają tworzenie drzew klasyfikacyjnych w przypadku nominalnej zmiennej zależnej. Algorytm C&RT pozwala tworzyć drzewa także w przypadku ciągłych zmiennych zależnych, podczas gdy CHAID i Exhaustive CHAID nie dawały pierwotnie takiej możliwości. W chwili obecnej rozszerzono te dwa ostatnie algorytmy (CHAID i Exhaustive CHAID) o możliwość rozpatrywania zmiennych zależnych typu porządkowego i ciągłego. Algorytm QUEST daje możliwość tworzenia drzew klasyfikacyjnych jedynie w przypadku nominalnej zmiennej zależnej.

Za pomocą każdego z wymienionych powyżej algorytmów możemy budować drzewo klasyfikacyjne, gdy rozpatrywane predyktory są nominalnymi, porządkowymi lub ciągłymi zmiennymi. Posługując się algorytmem CHAID lub Exhaustive CHAID, musimy jednak pamiętać, że predyktory ciągłe są przekształcane w predyktory porządkowe o (w przybliżeniu) jednakowo licznych kategoriach (domyślnie 10).

Algorytmy C&RT i QUEST umożliwiają utworzenie binarnych drzew klasyfikacyjnych. Tak więc na każdym poziomie drzewa zbiór obiektów dzieli się na dwa podzbiory. Dwa pozostałe algorytmy pozwalają utworzyć drzewa niebinarne, tj. mające przynajmniej jeden węzeł, z którego wychodzą więcej niż

[1] W nawiasach podajemy oryginalne nazwy angielskie używane w programie AnswerTree.

dwie krawędzie — zbiór obiektów jest wtedy dzielony na więcej niż dwa podzbiory na jednym poziomie.

Algorytmy różnią się zakresem wykorzystywania testów statystycznych w procesie budowy drzewa. Algorytmy CHAID, Exhaustive CHAID i QUEST używają testów statystycznych do doboru (selekcji) predyktorów, natomiast w algorytmie C&RT test statystyczny nie jest używany przy selekcji predyktorów.

W CHAID, Exhaustive CHAID testy statystyczne są wykorzystywane do podziału obiektów na klasy. W algorytmach C&RT i QUEST nie stosuje się testów statystycznych w procesie podziału obiektów.

W przeprowadzanych analizach dwa spośród wymienionych algorytmów wykorzystują informacje o prawdopodobieństwach *a priori*. Są to: C&RT i QUEST. W algorytmach CHAID i Exhaustive CHAID prawdopodobieństwa *a priori* nie są uwzględniane w analizie.

Wszystkie wymienione algorytmy wykorzystują podczas grupowania obiektów informacje o kosztach błędnych klasyfikacji (*misclassification costs*). Natomiast tylko algorytm C&RT używa kosztów błędnych klasyfikacji w budowaniu drzewa. W algorytmie QUEST symetryczne koszty błędnej klasyfikacji mogą być uwzględnione za pośrednictwem prawdopodobieństw *a priori*. Pod pojęciem kosztu błędnej klasyfikacji rozumie się ocenę straty („kary") spowodowanej przydzieleniem pewnemu obiektowi wartości y_i zmiennej zależnej, podczas gdy wynosi ona y_j. Formalnie biorąc, jest to nieujemna funkcja rzeczywista określona na iloczynie kartezjańskim $Y \times Y$ (gdzie Y — zbiór wartości zmiennej zależnej), tj. f: $Y \times Y \rightarrow R$, gdzie $f(y_i, y_j) = 0$, dla $i = j$. Funkcja jest symetryczna, gdy $f(y_i, y_j) = f(y_j, y_i)$. Ocena kosztu błędnej klasyfikacji dla podzbioru S_k wynosi

$$\min_c = \left\{ \sum_j f(y_i, y_j) \cdot p(p_j|k) \right\} \tag{8}$$

gdzie

$$p(j|k) = \frac{n_{ik}}{n_{.k}} \tag{9}$$

jest prawdopodobieństwem, że obiekt z podzbioru S_k ma wartość y_j.

Algorytmy różnią się sposobem traktowania brakujących danych. Algorytmy CHAID i Exhaustive CHAID traktują kategorie brakujących danych jako kategorie predyktorów. Natomiast C&RT i QUEST zastępują braki danych wartościami estymowanymi, np. średnimi, interpolacją danych.

3.2. Kryteria zakończenia budowy drzewa

Algorytm rekurencyjnego podziału działa tak długo, aż uzyskane podzbiory będą homogeniczne ze względu na zmienną zależną. Prowadzi to do tworzenia drzew o bardzo dużej złożoności (dużej głębokości i wielkości), w skrajnych przypadkach do podziału początkowego zbioru obiektów (n-elementowego) na jednoelementowe podzbiory $S_1, \ldots, S_n$. Ze względów praktycznych zastosowań godzimy się, aby uzyskane podzbiory nie były ściśle homogeniczne, ale bardziej liczne. Jedną ze stosowanych metod jest łączenie pod-

zbiorów w ten sposób, aby błąd klasyfikacji nie zwiększał się znacząco. Jest to tzw. przycinanie drzewa (*pruning*).

Proces budowania drzewa zostaje zakończony na poziomie danego węzła, jeżeli spełniony jest przynajmniej jeden z podanych poniżej warunków (kryteriów stopu):

- wszystkie obiekty węzła mają jednakową wartość zmiennej zależnej (predyktora);
- została osiągnięta maksymalna dopuszczalna liczba poziomów drzewa — maksymalna głębokość drzewa;
- liczba obiektów w węźle jest mniejsza od dopuszczalnej minimalnej liczby obiektów, która podlega podziałowi;
- podział węzła doprowadziłby do utworzenia węzła podrzędnego o liczebności mniejszej niż dopuszczalna minimalna liczba obiektów w węźle.

W przypadku algorytmu C&RT budowa drzewa może zostać zakończona także wówczas, gdy największy spadek zanieczyszczenia (*impurity*) byłby mniejszy niż podana przez użytkownika wymagana wartość minimalnego spadku zanieczyszczenia, a więc gdy zmiana (poprawa) jakości podziału jest zbyt mała. Na przykład

$$\max_t \left|\Delta H(S, x_t)\right| < g \tag{10}$$

gdzie g jest zadaną liczbą.

Wartości określające maksymalną liczbę poziomów drzewa, minimalną liczbę obiektów, która podlega podziałowi, minimalną liczbę obiektów w węźle wyznacza użytkownik metody.

3.3. Procedura przycinania drzew

Procedura jest stosowana w przypadku algorytmów C&RT i QUEST.

Celem przycinania drzewa (ang. *cost-complexity pruning*) jest znalezienie takiego drzewa, które byłoby jak najlepsze z punktu widzenia kosztu błędnej klasyfikacji, a zarazem nie byłoby nadmiernie rozbudowane. Zbyt rozbudowane drzewa są wrażliwe na obserwacje z próby, tj. tak silnie odwzorowują sytuację z próby, że zmniejsza to ich przydatność do klasyfikowania nowych obserwacji.

W C&RT i QUEST przycinanie drzew opiera się na zastosowaniu metody „jednego odchylenia standardowego" i polega na wyborze najprostszego drzewa spośród tych, dla których ocena kosztu błędnej klasyfikacji leży w granicach jednego odchylenia standardowego od najmniejszej uzyskanej oceny kosztu. Zwiększenie granic powyżej jedności prowadzi do wyboru prostszych drzew, a zmniejszenie: do wyboru drzew bardziej złożonych.

3.4. Walidacja i ocena ryzyka

Po zbudowaniu drzewa może zostać oszacowana jego wartość predykcyjna (*predictive value*). W przypadku wszystkich omówionych metod stosowane są takie same metody oceny ryzyka predykcji (błędu prognozowania).

W przypadku nominalnych i porządkowych zmiennych objaśnianych każdy węzeł przypisuje prognozowaną kategorię do wszystkich obserwacji należących do tej kategorii. Ocena ryzyka predykcji polega na wskazaniu, jaki jest udział przypadków niepoprawnie sklasyfikowanych.

W przypadku ciągłych zmiennych objaśnianych dla każdego węzła prognozowana jest jego wartość jako średnia wartość przypadków w węźle. Ocena ryzyka jest wariancją wewnątrzgrupową („wewnątrzwęzłową") wokół średniej, uśrednioną dla wszystkich węzłów — innymi słowy, jest to średni błąd kwadratowy w węzłach.

3.5. Ocena jakości utworzonego drzewa decyzyjnego

Jedną z metod oceny jakości utworzonego drzewa jest ocena udziału błędnie klasyfikowanych obiektów, które nie były wykorzystywane do budowy drzewa. Zbiór obiektów dzieli się wtedy na zbiór uczący, na którego podstawie następuje podział, i zbiór testowy obiektów niebiorących udziału w generowaniu drzewa, ale o znanej przynależności do klas.

Następnie tworzona jest tablica kontyngencji postaci:

	y_1	y_1	.	.	.	y_j	.	.	.	y_s	
$\hat{y}_1$	n_{11}	n_{12}	.	.	.					n_{1s}	$n_{1.}$
	n_{21}										$n_{2.}$
.											
.											
.											
$\hat{y}_2$						n_{ij}					$n_{i.}$
.											
.											
.											
$\hat{y}_r$	n_{r1}									n_{rs}	$n_{r.}$
	$n_{.1}$					$n_{.j}$				$n_{.s}$	N

(11)

gdzie:
y_i — prawdziwa wartość,
$\hat{y}_i$ — wartość przewidywana przez model,
n — liczba obiektów zbioru testowego,

wówczas:

$$e = \frac{\sum_{i,j=1,\, i \neq j}^{r} n_{ij}}{n} \tag{12}$$

jest miarą błędu klasyfikacji.

Ponieważ część posiadanych danych nie jest wykorzystywana podczas budowania drzewa, po zakończeniu budowy drzewa dane te są używane do oceny błędu niepoprawnej klasyfikacji. Takie postępowanie pomaga zidentyfikować drzewa, które dobrze odzwierciedlają specyfikę próbki danych użytych do ich budowy, ale nie przedstawiają prawidłowości odnoszących się do innych danych lub całej populacji.

Jeżeli nie mamy zbioru testowego, to błąd klasyfikacji należy estymować. Jedną ze stosowanych metod jest metoda sprawdzania krzyżowego (*cross-validation*):

- Dzieli się zbiór S na rozłączne i w przybliżeniu równoliczne podzbiory S_1, ..., S_s.
- Dla każdego $j = 1, ..., s$ buduje się model na podstawie zbioru obiektów S z wyłączeniem jednego podzbioru S_j, tzn. $S_j^* = S \setminus S_j$ i szacuje się błąd klasyfikacji e_j, traktując podzbiór S_j jako zbiór testowy.
- Następnie oblicza się wartość średnią $\bar{e} = \frac{\sum_{j=1}^{k} e_j}{k}$ tych błędów klasyfikacji.

Powyższa metoda gwarantuje uzyskanie nieobciążonych estymatorów błędu klasyfikacji. W programie C&RT można samodzielnie wybrać liczbę s określającą, na ile podzbiorów dzielimy zbiór obiektów.

Tak więc w metodzie sprawdzania krzyżowego podczas budowy drzewa wykorzystuje się wszystkie posiadane dane, ale są one dzielone na s oddzielnych grup — zbiorów (gdzie s jest określane przez użytkownika). Budowanych jest s drzew przy stosowaniu tych samych parametrów budowania drzew, jak w przypadku ocenianego drzewa. Przy budowaniu pierwszego drzewa używamy wszystkich zbiorów oprócz pierwszego, drugiego drzewa — wszystkich zbiorów oprócz drugiego, i tak dalej, aż każdy zbiór zostanie jeden raz wyłączony podczas budowy drzewa. Dla każdego zbudowanego drzewa dokonywana jest ocena ryzyka niepoprawnej klasyfikacji, aby ostatecznie przyjąć, że ryzyko niepoprawnej klasyfikacji dla drzewa jest średnią z s ocen ryzyka dla s drzew, ważoną liczbą przypadków w każdym używanym zbiorze danych.

4. Metody tworzenia drzew klasyfikacyjnych dostępne w programie AnswerTree i wykorzystywane do segmentacji przedsiębiorstw

4.1. Metoda CHAID i Exhaustive CHAID

Metoda CHAID (*Chi-squared Automatic Interaction Detection*) została pierwotnie opracowana jako metoda przeznaczona do tworzenia drzew klasyfikacyjnych tylko dla przypadku, gdy analizowane są zmienne dyskretne. Obecnie metoda została rozbudowana, tak że istnieją algorytmy CHAID dla przypadku nominalnych, porządkowych i ciągłych zmiennych objaśnianych. Zmienne diagnostyczne (objaśniające) ciągłe są zamieniane na zmienne dyskretne przed rozpoczęciem budowy drzewa.

Drzewo CHAID jest drzewem decyzyjnym, które jest konstruowane przez powtarzany iteracyjnie podział podzbiorów przestrzeni obiektów na dwa lub więcej wierzchołków — zwanych potomkami (dziećmi). Budowanie drzewa rozpoczyna się od podziału całego zbioru danych.

Aby wyznaczyć najlepszy podział dla wierzchołka, proponuje się [Kass, 1980] tworzenie wszystkich możliwych par kategorii zmiennej diagnostycznej (zbiór dopuszczalnych par jest zależny od typu analizowanej zmiennej diagnostycznej), takich, że nie ma statystycznie istotnej różnicy między elementami pary, rozpatrywanej w odniesieniu do zmiennej objaśnianej. Proces jest powtarzany aż do chwili, gdy żadna taka para nie zostanie znaleziona.

Opisane powyżej postępowanie jest przeprowadzane dla wszystkich zmiennych diagnostycznych. Zmienna diagnostyczna, która daje najlepszą predykcję, zostaje wybrana i wierzchołek jest dzielony na wierzchołki na kolejnym niższym poziomie drzewa. Proces jest powtarzany rekurencyjnie aż do zadziałania jednej z reguł zatrzymania budowy drzewa.

Opisane postępowanie nie daje gwarancji, że zostaną znalezione podziały, które są najlepsze w każdym wierzchołku drzewa. Jedynie pełne przeszukiwanie wszystkich możliwych podzbiorów kategorii może zapewnić najlepszy podział w każdym wierzchołku.

Algorytm pełnego przeszukiwania Exhaustive CHAID został zaproponowany przez Biggsa i innych [Biggs i inni, 1991]. Biggs zaproponował znajdywanie najlepszego podziału przez sukcesywne łączenie podobnych par, aż pozostanie pojedyncza para. Zbiór kategorii, dla którego istotność jest największa, jest przyjmowany jako najlepszy podział dla zmiennej diagnostycznej. Proces poszukiwania najlepszego podziału jest przeprowadzany dla wszystkich zmiennych diagnostycznych. Zmienna diagnostyczna, która daje najlepszą predykcję, zostaje wybrana i wierzchołek jest dzielony na wierzchołki leżące na niższym poziomie w strukturze drzewa. Proces jest powtarzany rekurencyjnie aż do uruchomienia jednej z przyjętych reguł zakończenia procedury budowania drzewa.

Wybierając CHAID, trzeba pamiętać o jej następujących cechach:

- możemy analizować jedną lub więcej zmiennych diagnostycznych. Zmienne diagnostyczne mogą być: ciągłe, porządkowe lub nominalne;
- możemy rozpatrywać jedną zmienną objaśnianą. Zmienna objaśniana może być: nominalna, porządkowa lub ciągła. Od rodzaju zmiennej objaśnianej zależy to, jaka metoda będzie stosowana do podziału wierzchołków: wybierany jest rodzaj algorytmu umożliwiający wykorzystanie informacji zawartej w zmiennej objaśnianej, w zależności od tego, czy jest nominalna, porządkowa, czy też ciągła;
- musimy przyjąć wartości parametrów algorytmu CHAID: poziomów istotności używanych w łączeniu i podziale oraz kryterium zatrzymania procesu podziału, wag dla obserwacji i częstości dla obserwacji.

Jest to metoda zalecana do stosowania gdy:

- chcemy tworzyć niebinarne podziały;
- model klasyfikacyjny, który jest tworzony za pomocą CHAID jest znacząco lepszy niż tworzony innymi metodami.

4.2. Metoda C&RT

Metoda C&RT (*Classification and Regression Trees*) [Breiman i inni, 1984] tworzy binarne drzewa decyzyjne. Drzewo C&RT jest konstruowane w procesie powtarzanego podziału podzbiorów zbioru danych. Tworzone są dwa wierzchołki na każdym poziomie struktury drzewa — rozpoczynając od całego zbioru danych. Do podziału zostaje wybrana zmienna diagnostyczna, która umożliwia największą redukcję tzw. zanieczyszczenia lub zróżnicowania. Celem jest utworzenie podzbiorów danych, które są możliwie jak najbardziej jednorodne ze względu na zmienną objaśnianą. Aby to osiągnąć, szuka się minimum funkcji mierzącej heterogeniczność zbioru. Funkcję tę nazywa się miarą zanieczyszczenia (ang. *impurity*). Są stosowane różne miary zanieczyszczenia, które powinny spełniać pewne aksjomaty (warunki) [Gatnar, 1998, s. 32]. Szuka się największej redukcji — zmniejszenia tej funkcji.

Aby dokonać podziału któregoś z wierzchołków, zmienne diagnostyczne są badane i oceniane, tak aby znaleźć najlepszy punkt podziału — w przypadku zmiennych ciągłych, lub grupowania kategorii — w przypadku zmiennych nominalnych i porządkowych. Ocena dokonywana jest na podstawie wskaźnika poprawy (zmniejszenia) „zanieczyszczenia". Zmienna diagnostyczna, która daje największą poprawę pod względem „zanieczyszczenia" (największe zmniejszenie „zanieczyszczenia"), zostaje wybrana do dokonania podziału wierzchołka na wierzchołki położone na niższym poziomie w strukturze drzewa. Proces jest powtarzany rekurencyjnie aż do uruchomienia jednej z przyjętych reguł zakończenia budowy drzewa.

Dla użytkownika metody C&RT istotne są jej następujące cechy:

- jedna lub więcej zmiennych diagnostycznych. Zmienne diagnostyczne mogą być: ciągłe, porządkowe lub nominalne;

- jedna zmienna objaśniana. Zmienna objaśniana może być: nominalna, porządkowa lub ciągła. Rodzaj zmiennej objaśnianej określa wybór metody, jaka zostanie użyta w modelu budowy drzewa;
- użytkownik określa parametry algorytmu C&RT: tzw. *priors* dla dyskretnej zmiennej objaśnianej (wartości startowe dla procesu estymacji; można je wybrać: na podstawie danych treningowych, równe dla wszystkich klas, dowolnie podane przez użytkownika), miarę zanieczyszczenia, koszty niepoprawnej klasyfikacji, wagi dla obserwacji i wagi dla częstości.
- C&RT powinniśmy używać, gdy:
- chcemy ograniczyć nasze drzewo do binarnych rozgałęzień;
- model klasyfikacji wytworzony za pomocą C&RT jest w stopniu znaczącym lepszy niż tworzony za pomocą innych metod;
- chcemy, aby macierze kosztów miały wpływ na wybór zmiennej;
- wymagane są koszty złożoności przycinania (*cost complexity pruning*) [Gatnar, 1998, s. 106] lub bezpośrednie reguły zatrzymania algorytmu generowania drzewa (tzw. reguły stopu — ang. *direct stopping rules*).

4.3. Metoda QUEST

Nazwa metody QUEST pochodzi od pierwszych liter słów: *Quick, Unbiased, Efficient, Statistical Tree*. Metoda po raz pierwszy została opisana w pracy Loh i Shih [1997]. Jest to algorytm klasyfikacji, który tworzy binarne drzewo decyzyjne — tak jak C&RT.

Przyczyną tworzenia drzew binarnych jest to, że pozwalają na zastosowanie technik, takich jak: przycinanie (*pruning*), zastosowanie wielu bezpośrednich reguł stopu (*direct stopping rules*), zastępczych rozgałęzień (*surrogate*).

Metoda ta różni się od poprzednio opisanych CHAID i C&RT tym, że w algorytmach CHAID i C&RT wybór zmiennej i wybór punktu podziału (rozgałęziania) odbywa się jednocześnie podczas budowania drzewa, podczas gdy w QUEST wybory te dokonywane są w odrębnych procedurach.

Metody pełnego przeszukiwania (*exhaustive search methods*), takie jak C&RT, mają tendencję do wyboru zmiennych o większej liczbie kategorii (zmienne jakościowe) lub wartości (zmienne ilościowe), które mogą dać w procesie budowy drzewa więcej rozgałęzień. To wprowadza obciążenie do modelu, które zmniejsza możliwość uogólniania wyników. Innym ograniczeniem C&RT jest nakład obliczeniowy na poszukiwanie rozgałęzień drzewa (dokonywanie podziału wierzchołków). Metoda QUEST została opracowana, aby przezwyciężyć wymienione wady algorytmu C&RT.

Dla każdego rozgałęzienia drzewa oceniana jest siła związku między każdą zmienną diagnostyczną i zmienną objaśnianą — w przypadku zmiennych diagnostycznych porządkowych i ciągłych: za pomocą testu F Fishera-Snedecora lub testu Levene'a, a w przypadku zmiennych diagnostycznych nominalnych: testu chi-kwadrat. Jeżeli zmienna objaśniana jest wielomianowa, do tworzenia podklas używana jest analiza skupień dzieląca zbiór na 2 podzbiory. Dobór zmiennych odbywa się na podstawie pewnej funkcji oceniającej ja-

kość podziału. Zmienna diagnostyczna, której związek ze zmienną objaśnianą jest najsilniejszy, zostaje wybrana do podziału (rozgałęziania). Aby znaleźć optymalny punkt rozgałęzienia, dla zmiennej diagnostycznej jest stosowana kwadratowa analiza dyskryminacji (*Quadratic Discriminant Analysis* — QDA). Proces jest powtarzany rekurencyjnie aż do zadziałania jednej z przyjętych reguł zakończenia.

Jeżeli chcemy zastosować QUEST, musimy pamiętać, że charakteryzuje się ona następującymi cechami:

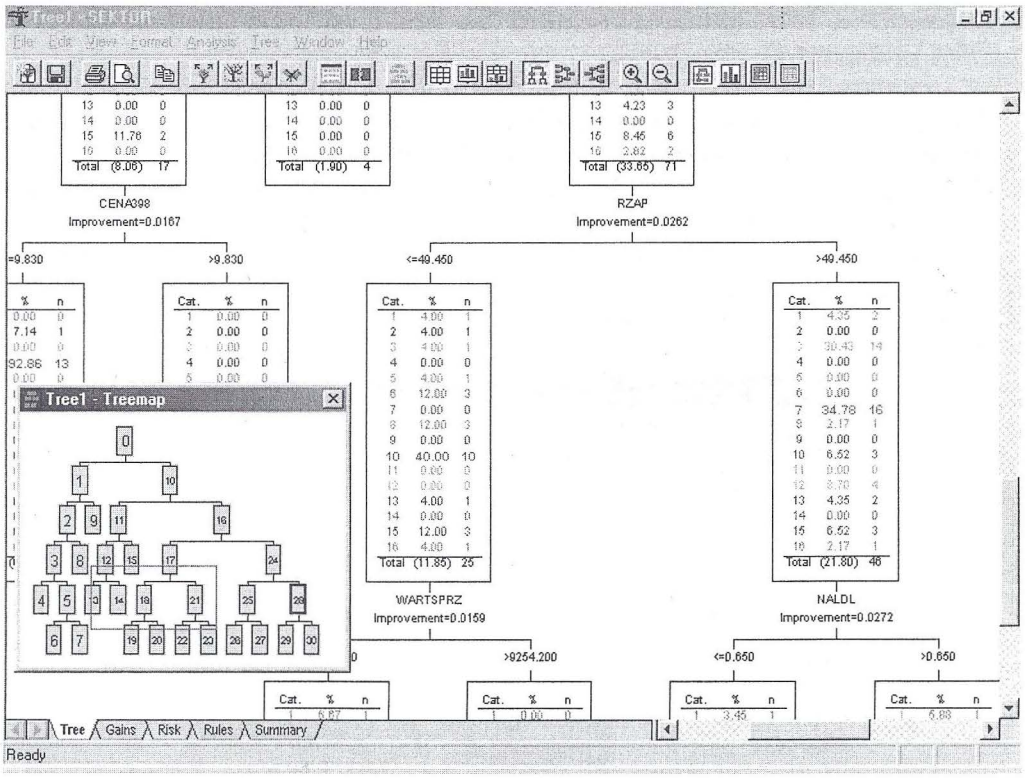
- możemy analizować jedną lub więcej zmiennych diagnostycznych. Zmienne diagnostyczne mogą być zmiennymi ciągłymi, porządkowymi lub nominalnymi;
- musimy wybrać jedną zmienną objaśnianą, która musi być zmienną nominalną;
- użytkownik musi ustalić parametry algorytmu, takie jak poziom istotności dla wyboru zmiennej, początkowe wartości w przypadku kategorycznych zmiennych docelowych, informację o względnej wartości każdej kategorii zmiennej, koszt błędnej klasyfikacji.

Powinniśmy używać QUEST:

- tylko w przypadku nominalnej zmiennej objaśnianej;
- jeżeli jest istotne uzyskanie nieobciążonego drzewa;
- jeżeli mamy duży lub złożony zbiór danych i potrzebujemy efektywnego algorytmu dla oszacowania drzewa;
- jeżeli chcemy zbudować drzewo posiadające tylko binarne rozgałęzienia;
- jeżeli model klasyfikacji wytworzony za pomocą QUEST jest w znaczącym stopniu (znacznie) lepszy niż wytworzony za pomocą innych metod;
- jeżeli chcemy, aby w przypadku brakujących wartości były stosowane rozgałęzienia zastępcze (*surrogate splits*);
- gdy nie potrzebujemy uwzględniać wag dla obserwacji;
- jeżeli chcemy stosować zdefiniowane przez nas reguły zatrzymania budowy drzewa i przycinanie drzewa;
- gdy macierz kosztów nie jest bezpośrednio włączona w proces tworzenia drzewa.

5. Przykłady przeprowadzania segmentacji przedsiębiorstw za pomocą drzew klasyfikacyjnych

Na rys. 2. przedstawiono ekran z widocznym fragmentem drzewa decyzyjnego podziału przedsiębiorstw według przynależności do różnych sektorów: 1 — media, 2 — przemysł spożywczy, 3 — przemysł elektromaszynowy, 4 — handel, 5 — informatyka i telekomunikacja, 6 — usługi komunalne, 7 — przemysł lekki, 8 — przemysł materiałów budowlanych, 9 — usługi finansowe, 10 — budownictwo, 11 — pozostałe usługi, 12 — przemysł chemiczny, 13 — przemysł motoryzacyjny, 14 — konglomerat, 15 — przemysł metalowy, 16 — przemysł drzewny i papierniczy.



Rys. 2.

Ekran programu AnswerTree z widocznym fragmentem drzewa decyzyjnego wygenerowanego za pomocą metody C&RT

Źródło: program AnswerTree.

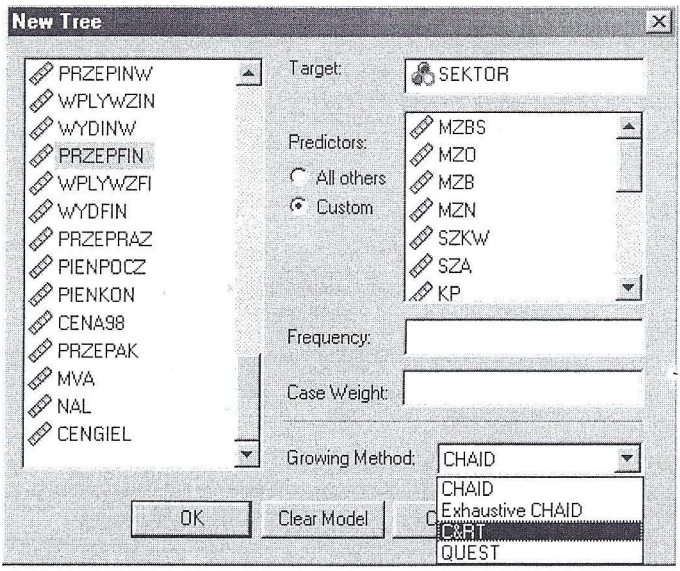
Zmienną objaśnianą jest zmienna określająca przynależność przedsiębiorstw do sektorów, przyjmująca wartości od 1 do 16. Jest to zmienna nominalna o nazwie SEKTOR. Pozostałe zmienne przedstawiające wielkości i wskaźniki finansowe i majątkowe są zmiennymi objaśniającymi. Są to zmienne ilościowe ciągłe. Jako metodę generowania drzewa wybrano metodę C&RT (por. rys. 3.).

Na rys. 2. widoczny jest węzeł drzewa, w którym przedsiębiorstwa są dzielone według rotacji zapasów (zmienna RZAP): RZAP $\leq$ 49,45 lub RZAP > 49,45. Ponieważ zastosowano metodę C&RT, podawana jest wartość określająca stopień redukcji zanieczyszczenia *Improvement*. W metodzie C&RT predyktor z największym zmniejszeniem zanieczyszczenia jest wybierany do kolejnego podziału w węźle drzewa. W analizowanym węźle dokonano podziału 71 przedsiębiorstw. Spośród przedsiębiorstw, które mają wartości RZAP $\leq$ 49,45 (jest ich 25), najwięcej należy do sektora 10 — budownictwo (10 spośród 25, tj. 40% przedsiębiorstw). Najwięcej przedsiębiorstw o wartoś-

ciach RZAP > 49,45 należy do sektora 7 — przemysł lekki (16 spośród 46, tj. 34,78% przedsiębiorstw).

Z przedstawionego na rys. 2. fragmentu drzewa możemy odczytać, że do dalszego podziału wykorzystywane były wartość sprzedanych towarów i materiałów (WARTSPRZ) i należności długoterminowe (NALDL).

Widoczne jest okno przedstawiające tzw. mapę — plan całego drzewa (*treemap*) i ułatwiające w programie AnswerTree przemieszczanie się do wybranego fragmentu drzewa. Drzewo podziału przedsiębiorstw na sektory ma 5 poziomów (szósty poziom tworzy korzeń drzewa — *root nodel*) i składa się z 30 węzłów oraz węzła korzenia.



Rys. 3.

Wprowadzanie zmiennych do modelu i wybór metody generowania drzewa

Źródło: program AnswerTree.

Sposób prezentacji wyników w węzłach drzewa zależy od skali pomiaru zmiennej objaśnianej — jak ilustrują to rys. 4a i 4b.

W przypadku zmiennych nominalnych lub porządkowych dla każdej kategorii podawana jest liczebność i procentowy udział w ogólnej liczebności. W przypadku zmiennych ciągłych podawana jest średnia wartość, odchylenie standardowe, liczebność, a także wartość prognozowana średniej. Jeżeli prognozowanie nie było przeprowadzane, to wartość ta jest równa wartości średniej. Na rys. 4. przedstawiono węzeł drzewa dla przypadku, gdy zmienna objaśniana jest zmienną nominalną (zmienna SEKTOR: rys. 4a) oraz węzeł drzewa dla przypadku, gdy zmienna objaśniana jest zmienną ciągłą (zmienna WKAK — wartość księgowa na jedną akcję: rys. 4b).

a)

SEKTOR

Cat.	%	n
1	2.84	6
2	12.32	26
3	11.37	24
4	7.11	15
5	6.64	14
6	1.90	4
7	9.00	19
8	6.16	13
9	0.95	2
10	**18.96**	**40**
11	2.84	6
12	5.69	12
13	3.32	7
14	0.47	1
15	5.21	11
16	5.21	11
Total (100.00)		211

b)

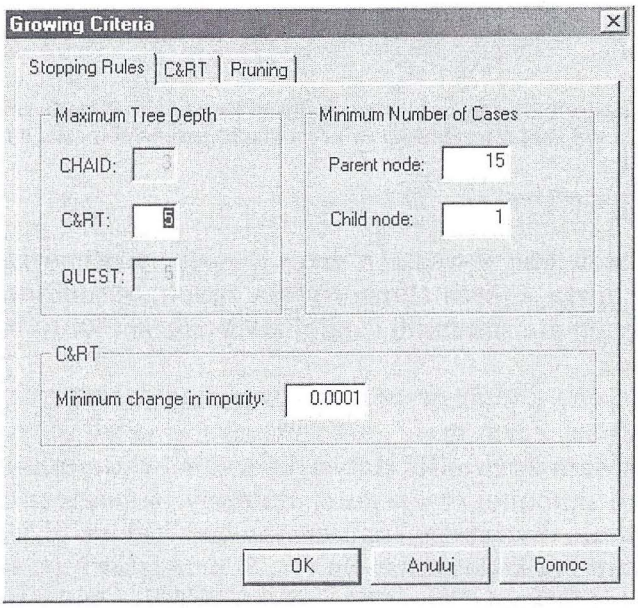
WKAK

Mean	18.88001
Std.Dev	16.60387
n	211 (100.00%)
Predicted	18.88001

Rys. 4.

Prezentacja wyników w węźle drzewa w zależności od skali pomiaru zmiennej (a) — zmienna nominalna lub porządkowa, (b) zmienna ilościowa ciągła

Źródło: program AnswerTree.



Rys. 5.

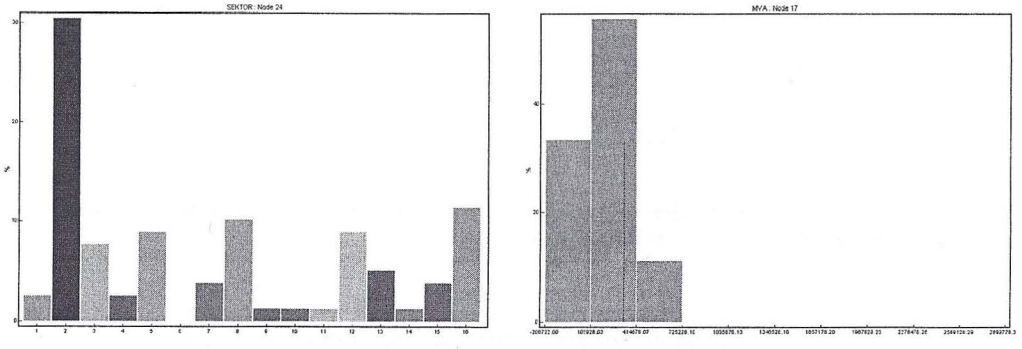
Definiowanie reguł zatrzymania procedury generowania drzewa w przypadku zastosowania metody C&RT

Źródło: program AnswerTree.

Jeżeli tworzymy drzewa za pomocą algorytmów dostępnych w programie AnswerTree, to musimy określić warunki zakończenia ich działania (*Stopping Rules*). W przypadku zastosowania metody C&RT — jak ilustruje to rys. 5. — musimy określić:

- maksymalną głębokość drzewa (*Maximum Tree Depth*),
- minimalną liczbę obserwacji w węźle dzielonym (*Minimum Number of Cases Parent Node*),
- minimalną liczbę obserwacji w powstającym węźle (*Minimum Number of Cases Child Node*),
- minimalną redukcję zanieczyszczenia (*Minimum Change in Impurity*).

a b



Rys. 6.

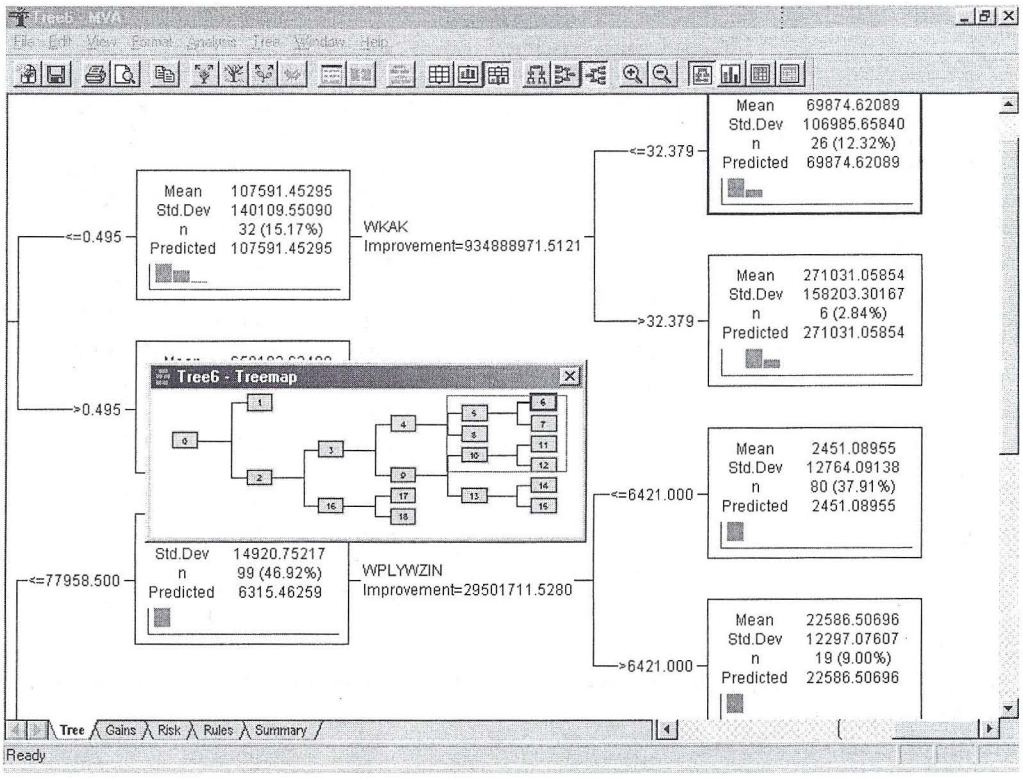
Graficzna analiza danych: a — wykres przynależności przedsiębiorstw do sektorów dla analizy 24. węzła drzewa, b — wykres liczebności (w procentach) przedsiębiorstw o różnych wielkościach MVA

Źródło: program AnswerTree.

Dane przedstawiane w węzłach drzewa możemy zilustrować wykresami. Rys. 6a przedstawia wykres słupkowy dla zmiennej nominalnej SEKTOR, a rys. 6b histogram dla zmiennej ciągłej MVA (*Market Value Added*: rynkowa wartość dodana).

Wyniki obliczeń możemy przedstawić w postaci tabelarycznej i graficznej na jednym rysunku — por. rys. 7., ale wykresy nie są wówczas zbyt czytelne.

Program pozwala wyświetlić statystykę wyników (*Gain Summary*) dla wybranej kategorii zmiennej objaśnianej. Załóżmy, że interesują nas przedsiębiorstwa przemysłu drzewnego i papierniczego — sektor 16. według przyjętego przez nas systemu kodowania. Na rys. 8. przedstawiono tabelę wyników dla zmiennej objaśnianej (*Target Variable*) SEKTOR i kategorii (*Target Category*) 16. — przedsiębiorstwa przemysłu drzewnego i papierniczego. Kolejne kolumny tabeli przedstawiają: numer węzła (*Node*) — np. 29, łączną liczbę przedsiębiorstw w tym węźle (*Node: n*) — 53, udział w procentach przedsiębiorstw węzła w całkowitej liczbie analizowanych przedsiębiorstw (*Node: %*)



Rys. 7.

Wyniki w postaci tabelarycznej i graficznej na jednym ekranie (na przykładzie drzewa generowanego dla zmiennej objaśnianej MVA (rynkowa wartość dodana)

Źródło: program AnswerTree.

— 53/211 = 0,2512 = 25,12%, liczbę przedsiębiorstw należących do badanej kategorii i analizowanego węzła (*Resp: n*) — 7 przedsiębiorstw należących do 16. sektora w węźle 29. i ich udział w łącznej istniejącej liczbie przedsiębiorstw tego sektora — w naszych badaniach było 11 przedsiębiorstw należących do 16. sektora (*Resp: %*) — 7/11 = 0,6364 = 63,64%. Kolumna *Gain (%)* pokazuje, jaki w danym węźle jest udział przedsiębiorstw należących do 16. sektora (*Target Category*) w łącznej liczbie przedsiębiorstw z tego węzła: np. dla węzła 29. jest to 7/53 = 0,1320755 = 13,20755%. Ostatnia kolumna *Index (%)* to wielkości z kolumny *Gain (%)* podzielone przez obliczony dla korzenia drzewa procentowy udział przedsiębiorstw należących do 16. sektora (u nas *Target Category*) w łącznej liczbie wszystkich analizowanych przedsiębiorstw. Dla węzła 29. wynosi: 253,34477, gdzie procentowy udział przedsiębiorstw należących do 16. sektora możemy odczytać z rys. 3a — wynosi 5,21%.

Gain Summary						
Target Variable: SEKTOR Target Category: 16						
	Statistics					
Node	Node: n	Node: %	Resp: n	Resp: %	Gain (%)	Index (%)
29	53	25.12	7	63.64	13.20755	253.34477
26	17	8.06	2	18.18	11.76471	225.66845
19	15	7.11	1	9.09	6.66667	127.87879
23	17	8.06	1	9.09	5.88235	112.83422
22	29	13.74	0	0.00	0.00000	0.00000
6	27	12.80	0	0.00	0.00000	0.00000
13	14	6.64	0	0.00	0.00000	0.00000
20	10	4.74	0	0.00	0.00000	0.00000
9	8	3.79	0	0.00	0.00000	0.00000
30	5	2.37	0	0.00	0.00000	0.00000
15	4	1.90	0	0.00	0.00000	0.00000
27	4	1.90	0	0.00	0.00000	0.00000
14	3	1.42	0	0.00	0.00000	0.00000
8	3	1.42	0	0.00	0.00000	0.00000
7	1	0.47	0	0.00	0.00000	0.00000
4	1	0.47	0	0.00	0.00000	0.00000

Rys. 8.

Analiza statystyczna węzłów drzewa przy założeniu kategorii docelowej: przedsiębiorstwa z 16. sektora

Źródło: program AnswerTree.

Drzewo klasyfikacyjne może nam posłużyć do odczytania reguł segmentacji przedsiębiorstw. Program generuje te reguły w postaci zdań logicznych (implikacji). Jedną z reguł analizy przynależności przedsiębiorstw do sektorów przedstawiono na rys. 9.

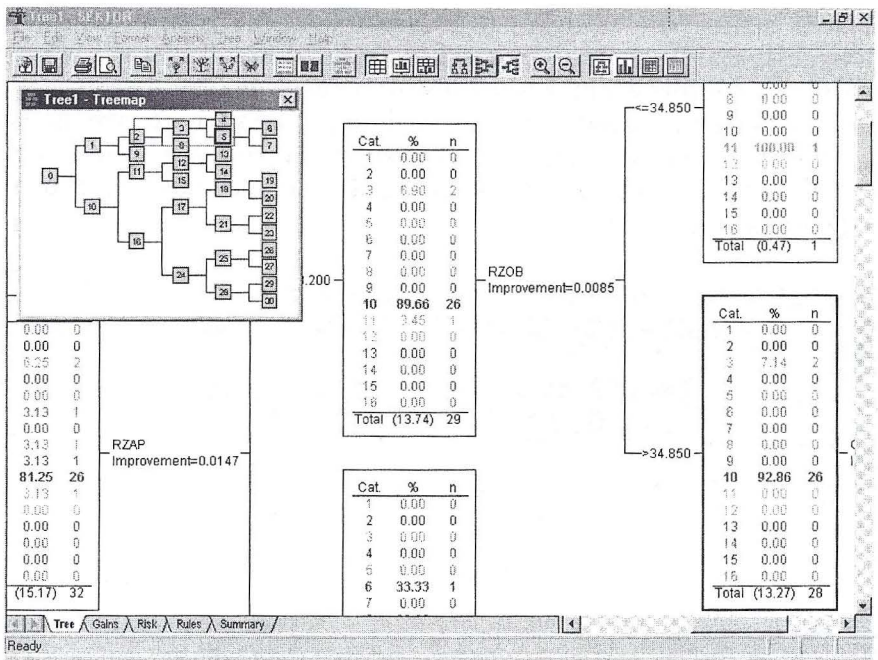
```
/* Node 5 */
if kosztspr <= 76.4 and mzo <= 0.11 and rzap <= 48.2 and rzob > 34.85
then
    node = 5
    prediction = 10
    probability = 0.929
```

Rys. 9.

Przykład reguły odczytanej z drzewa klasyfikacyjnego przynależności przedsiębiorstw do różnych sektorów (oznaczenia: kosztspr — koszty sprzedanych produktów, towarów i materiałów, mzo — marża zysku operacyjnego, rzap — rotacja zapasów, rzob — rotacja zobowiązań)

Źródło: program AnswerTree.

Przedstawiona powyżej reguła odpowiada 5. węzłowi drzewa — rys. 10.



Rys. 10.

Fragment drzewa decyzyjnego z widocznym 5. węzłem, dla którego przedstawiono regułę na rys. 9.

Źródło: program AnswerTree.

Do analizy poprawności klasyfikacji zostaje utworzona tablica kontyngencji (por. wzór 11), tak jak przedstawiono to na rys. 11. dla przypadku przydzielania przedsiębiorstw do sektorów.

Misclassification Matrix

Predicted Category \ Actual Category	1	2	3	4	5	6	7	8	9
1	2	0	0	0	0	0	0	0	0
2	2	22	4	0	2	0	3	8	1
3	0	0	12	0	0	0	6	0	0
4	0	3	1	13	2	0	0	0	0
5	2	0	3	0	9	0	0	0	0
6	0	0	0	0	0	2	1	1	1
7	0	1	0	2	1	0	9	4	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	4	0	0	2	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0
total	6	26	24	15	14	4	19	13	2

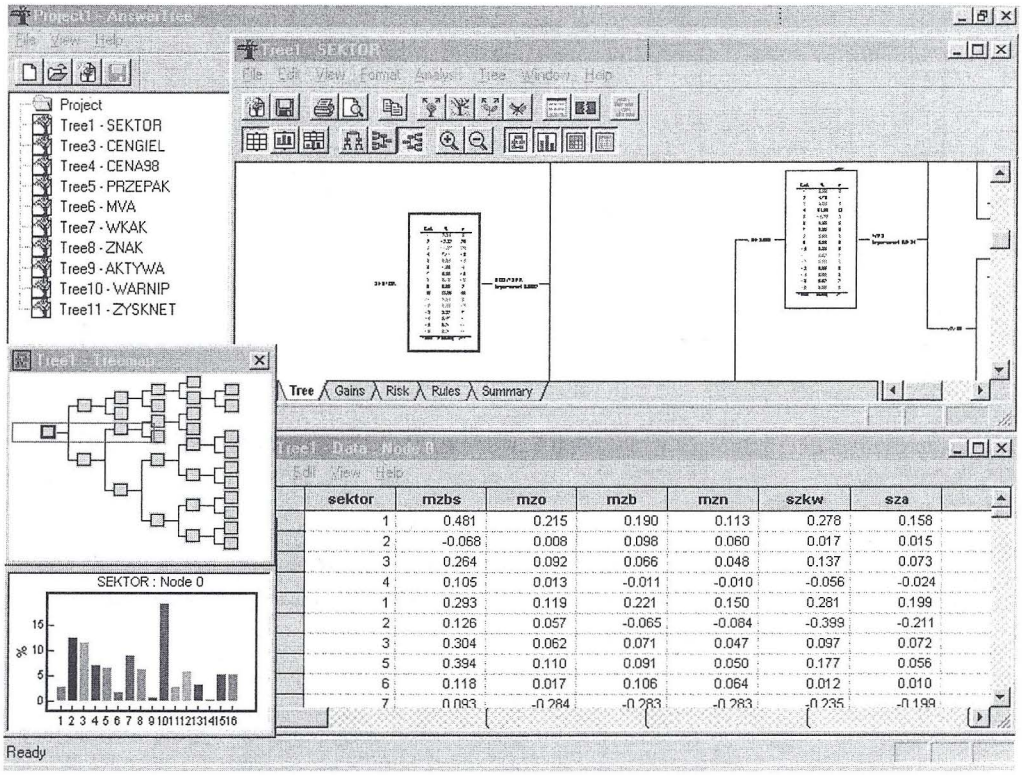
Rys. 11.

Analiza błędów klasyfikacji — fragment tabeli dla przykładu przydzielania przedsiębiorstw do jednego z 16 sektorów

Źródło: program AnswerTree.

W tabeli zestawiane są liczby przedsiębiorstw faktycznie należących do określonych sektorów (*Actual Category*) z liczbami przedsiębiorstw przydzielanymi do tych sektorów przez wygenerowane drzewo (*Predicted Category*).

Jeżeli korzystamy z programu AnswerTree, to analizę otrzymanego drzewa możemy przeprowadzać, przedstawiając wiele różnych wyników (okien) na jednym ekranie — rys. 12., a także sporządzając tekstowy raport opisujący model generowania drzewa.



sektor	mzbs	mzo	mzb	mzn	szkw	sza
1	0.481	0.215	0.190	0.113	0.278	0.158
2	-0.068	0.008	0.098	0.060	0.017	0.015
3	0.264	0.092	0.066	0.048	0.137	0.073
4	0.105	0.013	-0.011	-0.010	-0.056	-0.024
1	0.293	0.119	0.221	0.150	0.281	0.199
2	0.126	0.057	-0.065	-0.084	-0.399	-0.211
3	0.304	0.062	0.071	0.047	0.097	0.072
5	0.394	0.110	0.091	0.050	0.177	0.056
6	0.118	0.017	0.106	0.064	0.012	0.010
7	0.093	-0.284	-0.283	-0.283	-0.235	-0.199

Rys. 12.

Posługiwanie się wieloma oknami na jednym ekranie

Źródło: program AnswerTree.

W celu oceny jakości utworzonego drzewa klasyfikacyjnego część danych możemy wydzielić jako dane testowe i przeprowadzić za ich pomocą testowanie po wygenerowaniu drzewa. Możemy też przeprowadzić testowanie krzyżowe (*cross-validation*), dzieląc zbiór danych na ustaloną liczbę podzbiorów. Możliwości dodatkowych analiz, nieprzedstawionych w analizowanym przykładzie, są duże. Możemy przykładowo wprowadzić wagi lub prawdopodobieństwa *a priori* (*prior probabilities*) dla różnych kategorii zmiennej zależnej, przeprowadzić analizę ryzyka opartego na analizie wariancji, zróżnicować

wagi — nadać różny koszt — różnych przypadków niepoprawnych klasyfikacji przez drzewo.

Podsumowanie

Próby wykorzystania drzew klasyfikacyjnych do segmentacji przedsiębiorstw wskazują, że są one dobrym narzędziem i ułatwiają przeprowadzanie analiz w wielu różnych przekrojach. Ich główne zalety w stosunku do klasycznych (tj. parametrycznych) metod (np. analizy regresji, analizy dyskryminacji) są następujące:

- unika się konieczności weryfikowania założeń dotyczących rozkładów zmiennych objaśniających;
- w modelu mogą występować jednocześnie zmienne jakościowe i ilościowe;
- metody wykazują odporność (małą wrażliwość) na występowanie wartości nietypowych — odstających (*outliers*) dla zmiennych objaśniających;
- wykazują odporność na występowanie brakujących wartości obserwowanych zmiennych;
- dobór zmiennych objaśniających jest dokonywany automatycznie podczas działania algorytmu.

Drzewo klasyfikacyjne utworzone na podstawie przykładów można wykorzystać do klasyfikacji nowych obiektów — takich, które nie były wykorzystywane do budowy drzewa. Rozpoczynając od korzenia drzewa, przechodzimy od wierzchołka do wierzchołka wzdłuż krawędzi drzewa, które odpowiadają wartościom cech klasyfikowanego obiektu — przedsiębiorstwa. Aby zaklasyfikować przedsiębiorstwo do określonej klasy, zazwyczaj nie musimy znać wszystkich jego cech, ponieważ najczęściej tylko niektóre cechy decydują o przynależności do określonej klasy. Jest to niewątpliwie zaletą drzew decyzyjnych w ich zastosowaniach do klasyfikowania „nowych" obiektów.

Problemy czasem stwarza duża złożoność drzewa, a także możliwość różnej interpretacji uzyskanych wyników. Nie ma także żadnych wskazówek co do wyboru modelu, który najlepiej pasowałby do analizowanego problemu. Ustalenia takie jak chociażby wybór metody generowania drzewa, liczby poziomów drzewa, reguł zatrzymania procedury generującej drzewo są podejmowane dosyć arbitralnie. Przydatne jest przeprowadzanie wielu różnych eksperymentów, przyjmując różne modele i założenia, co znakomicie ułatwia posiadanie odpowiedniego oprogramowania.

Literatura

AnswerTree. User's Guide, 1998, SPSS Inc., Chicago.

Biggs D., de Ville B., Suen E., 1991, *A method of choosing multiway partitions for classification and decision trees*, „Journal of Applied Statistics" nr 18, s. 49–62.

Breiman L., Friedman J., Olshen R., Stone C., 1984, *Classification and Regression Trees*, CRC Press, London.

Gatnar E., 2001, *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.
Gatnar E., 1998, *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.
http://www.spss.pl
Kass G., 1980, *An exploratory technique for investigating large quantities of categorical data*, „Applied Statistics" 29:2, s. 119–127.
Kovalerchuk B., Vityaev E., 2000, *Data Mining in Finance. Advances in Relational and Hybrid Methods*, Kluwer.
Lasek M., Pęczkowski M., 1991, *Knowledge Acquistition Method of Expert System for Analysis of Financial Standing of an Enterprise*, w: Alty J. L., Mikulich L. I. (wyd.), *Industrial Applications of Artificial Intelligence*, North-Holland, s. 81–86.
Loh W., Shih Y., 1997, *Split selection methods for classification trees*, Statistica Sinics.
Shi Y., 2000, *Data Mining*, w: Zeleny M., *The IEBM Handbook of Information Technology in Business*, Business Press, Thomson Learning, s. 490–495.
Witten J. H., Frank E., 2000, *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*, Academic Press, Morgan Kaufmann Publishers.
Zakrzewicz M., 2001, *Eksploracja danych (Data Mining) — zbiór technik automatycznego odkrywania nietrywialnych zależności i schematów w dużych zbiorach danych*, http://www.cs.put.poznan.pl/mzakrzewicz.

Załącznik. Struktura zbioru „Przedsiębiorstwa" wykorzystywanego do generowania drzew klasyfikacyjnych

Nazwa zmiennej	Opis zmiennej	Źródło lub sposób obliczania	Kategoria zmiennej
mzbs	Marża zysku brutto ze sprzedaży	[przychody ze sprzedaży netto — koszt wytworzenia produkcji sprzedanej]/przychody ze sprzedaży netto	ciągła
mzo	Marża zysku operacyjnego	zysk operacyjny/przychody ze sprzedaży netto	ciągła
mzb	Marża zysku brutto	zysk brutto/przychody ze sprzedaży netto	ciągła
mzn	Marża zysku netto	zysk netto/przychody ze sprzedaży netto	ciągła
szkw	Stopa zwrotu z kapitału własnego	zysk netto/średnia wartość kapitałów własnych	ciągła
sza	Stopa zwrotu z aktywów	zysk netto/średnia wartość aktywów	ciągła
kp	Kapitał pracujący	majątek obrotowy — zobowiązania krótkoterminowe	ciągła
wpb	Wskaźnik płynności bieżącej	majątek obrotowy/zobowiązania krótkoterminowe	ciągła
wps	Wskaźnik płynności szybkiej	[majątek obrotowy — zapasy]/zobowiązania krótkoterminowe	ciągła
wpp	Wskaźnik podwyższonej płynności	[majątek obrotowy — zapasy — należności]/zobowiązania krótkoterminowe	ciągła

Nazwa zmiennej	Opis zmiennej	Źródło lub sposób obliczania	Kategoria zmiennej
rmob	Rotacja majątku obrotowego w dniach	[średnia wartość majątku obrotowego/ przychody ze sprzedaży netto] × liczba dni	ciągła
rnal	Rotacja należności w dniach	[średnia wartość należności/przychody ze sprzedaży netto] × liczba dni	ciągła
rzap	Rotacja zapasów w dniach	[średnia wartość zapasów/koszt wytworzenia produkcji sprzedanej] × liczba dni	ciągła
cop	Cykl operacyjny w dniach	rotacja należności + rotacja zapasów	ciągła
rzob	Rotacja zobowiązań w dniach	[średnia wartość zobowiązań krótkoterminowych/koszt wytworzenia produkcji sprzedanej] × liczba dni	ciągła
ckg	Cykl konwersji gotówki w dniach	cykl operacyjny — rotacja zobowiązań Stopa zadłużenia (szad): zobowiązania/ aktywa	ciągła
wsp	Wskaźnik struktury pasywów	zobowiązania/kapitał własny	ciągła
wdf	Wskaźnik dźwigni finansowej	aktywa/kapitał własny	ciągła
aktywa	Aktywa	z bilansu	ciągła
warnip	Wartości niematerialne i prawne	z bilansu	ciągła
naldl	Należności długoterminowe	z bilansu	ciągła
zapasy	Zapasy	z bilansu	ciągła
nalkr	Należności krótkoterminowe	z bilansu	ciągła
spien	Środki pieniężne	z bilansu	ciągła
pasywa	Pasywa	z bilansu	ciągła
kw	Kapitał własny	z bilansu	ciągła
kakc	Kapitał akcyjny	z bilansu	ciągła
kzap	Kapitał zapasowy	z bilansu	ciągła
krez	Kapitał rezerwowy z aktualizacji wyceny	z bilansu	ciągła
zysknet	Zysk (strata) netto	z bilansu	ciągła
zob	Zobowiązania	z bilansu	ciągła
zobdl	Zobowiązania długoterminowe	z bilansu	ciągła
zobkr	Zobowiązania krótkoterminowe	z bilansu	ciągła
przynet	Przychody netto ze sprzedaży produktów, towarów i materiałów	z rachunku zysków i strat	ciągła
kosztsp	Koszty sprzedanych produktów, towarów i materiałów	z rachunku zysków i strat	ciągła
koszt wytw	Koszty wytworzenia sprzedanych produktów	z rachunku zysków i strat	ciągła
wartsprz	Wartość sprzedanych towarów i materiałów	z rachunku zysków i strat	ciągła
zyskbrus	Zysk (strata) brutto na sprzedaży	z rachunku zysków i strat	ciągła
kosztsprzed	Koszty sprzedaży	z rachunku zysków i strat	ciągła

Nazwa zmiennej	Opis zmiennej	Źródło lub sposób obliczania	Kategoria zmiennej
kosztzarz	Koszty ogólnego zarządu	z rachunku zysków i strat	ciągła
zysksprz	Zysk (strata) na sprzedaży	z rachunku zysków i strat	ciągła
zyskdo	Zysk (strata) na działalności operacyjnej	z rachunku zysków i strat	ciągła
zyskbru	Zysk (strata) brutto	z rachunku zysków i strat	ciągła
zysknet	Zysk (strata) netto	z rachunku zysków i strat	ciągła
przepopnet	Przepływy operacyjne netto	z rachunku przepływu środków pieniężnych	ciągła
amort	Amortyzacja	z rachunku przepływu środków pieniężnych	ciągła
poddoch	Podatek dochodowy (wskazany w rachunku zysków i strat)	z rachunku przepływu środków pieniężnych	ciągła
poddochzap	Podatek dochodowy zapłacony	z rachunku przepływu środków pieniężnych	ciągła
przepinwnet	Przepływy inwestycyjne netto	z rachunku przepływu środków pieniężnych	ciągła
wplywzinw	Wpływy z działalności inwestycyjnej	z rachunku przepływu środków pieniężnych	ciągła
wydinw	Wydatki z tytułu działalności inwestycyjnej	z rachunku przepływu środków pieniężnych	ciągła
przepfinnet	Przepływy finansowe netto	z rachunku przepływu środków pieniężnych	ciągła
wplywzfin	Wpływy z działalności finansowej	z rachunku przepływu środków pieniężnych	ciągła
wydfin	Wydatki z działalności finansowej	z rachunku przepływu środków pieniężnych	ciągła
przeprazem	Przepływy pieniężne razem	z rachunku przepływu środków pieniężnych	ciągła
pienpocz	Środki pieniężne na początek okresu	z rachunku przepływu środków pieniężnych	ciągła
pienkon	Środki pieniężne na koniec okresu	z rachunku przepływu środków pieniężnych	ciągła
rokrej	Rok rejestracji	Wyniki finansowe spółek giełdowych (Notoria Serwis)	porządkowa
zatr	Zatrudnienie	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
dywid	Dywidenda	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
akcje	Średnia liczba akcji (tys. sztuk)	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
wkak	Wartość księgowa na jedną akcję	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
znak	Zysk netto na jedną akcję	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
cenakrr	średni kurs akcji w k kwartale roku rr (k oznacza numer kwartału, a rr — rok), np. cena198 — średni kurs akcji w pierwszym kwartale 1998 r.	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
cenarr	średni roczny kurs akcji w roku rr, np. cena98 — średni kurs akcji w roku 1998	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła

Nazwa zmiennej	Opis zmiennej	Źródło lub sposób obliczania	Kategoria zmiennej
przepak	Przepływy pieniężne przypadające na jedną akcję	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
mva: market value added	Rynkowa wartość dodana	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
nal	Należności razem: razem należności długoterminowe i należności krótkoterminowe	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
cengiel	Cena giełdowa przedsiębiorstwa: średni kurs akcji pomnożony przez ich liczbę	Wyniki finansowe spółek giełdowych (Notoria Serwis)	ciągła
sektor	Przynależność przedsiębiorstwa do sektora	Wyniki finansowe spółek giełdowych (Notoria Serwis)	nominalna

Abstract **Segmentation of Firms by Means of Classification Trees**

The objective of the paper was to present the utility and applicability of the method of generating classification trees for the purposes of segmentation of firms by their economic standing, i.e. their financial and assets condition. The method of classification tree generation belongs to the group of the "data mining" methods that permit to find out, basing on large data sets, the relationships and links among data. Variables used to classify the firms were financial and assets indices, indices of the firms' position in the capital market, as well as values from financial statements (balance sheet, profit and loss account, cash flow account). Different algorithms were used to generate the classification trees: Chaid, Exhaustion Chaid, C&RT, Quest. In the paper, examples were presented of applying the classification trees to segmentation of firms. Advantages and drawbacks of performing the segmentation of firms by means of the classification trees were discussed.