



DRZEWIA KLASYFIKACYJNE W BADANIACH SATYSFAKCJI I LOJALNOŚCI KLIENTÓW

Mariusz Łapczyński

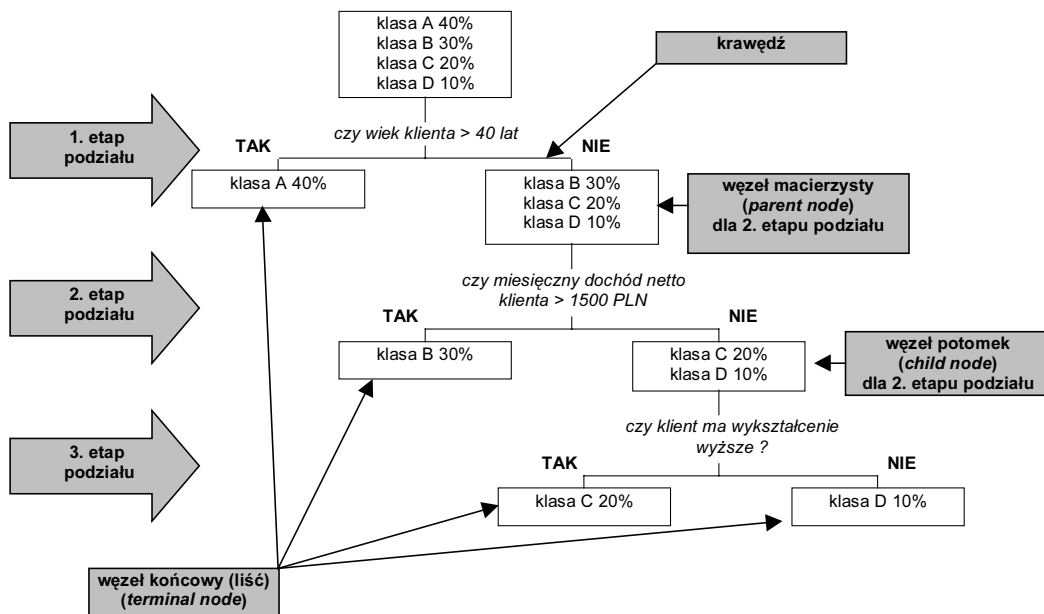
Akademia Ekonomiczna w Krakowie, Katedra Analizy Rynku i Badań Marketingowych

Wprowadzenie

Drzewa klasyfikacyjne i regresyjne to narzędzia *data mining* służące do budowy modeli predykcyjnych i deskryptywnych. Z drzewami klasyfikacyjnymi mamy do czynienia wtedy, gdy zmienna zależna jest wyrażona na skali nominalnej lub porządkowej, natomiast z drzewami regresyjnymi wtedy, gdy poziom pomiaru tej zmiennej jest co najmniej przedziałowy. Celem budowy modelu predykcyjnego jest predykcja jakościowa lub ilościowa zjawiska, zaś celem budowy modelu deskryptywnego jest opis i prezentacja wzorców w badanej zbiorowości.

Drzewo jest graficznym modelem powstałym w wyniku rekurencyjnego podziału zbioru obserwacji A na n rozłącznych podzbiorów $A_1, A_2, A_3, \dots, A_n$. Celem budowy modelu jest uzyskanie podzbiorów maksymalnie jednorodnych z punktu widzenia wartości zmiennej zależnej. Jest to proces wieloetapowy, który w każdym kolejnym kroku może wykorzystywać inną zmienną niezależną. Na każdym etapie analizuje się bowiem wszystkie predyktory i wybiera ten, który zapewnia najlepszy podział węzła, czyli wydziela najbardziej homogeniczne podzbiory.

Przykład drzewa klasyfikacyjnego przedstawiono na rys. 1. Początek każdego drzewa stanowi cały zbiór obserwacji, który jest dzielony na 2 lub więcej podzbiorów. W pierwszym przypadku mówi się o drzewach binarnych, a w drugim o drzewach dowolnych. Dzielony zbiór nosi nazwę węzła macierzystego (ang. *parent node*), natomiast wydzielone podzbiory – nazwę węzłów potomków (ang. *child nodes*). W kolejnym etapie podziału węzeł potomek, który jest dalej dzielony, staje się węzłem macierzystym dla 2. etapu, zaś węzeł, który pozostaje bez zmian, staje się węzłem końcowym, nazywanym liściem. Wielkość drzewa to liczba liści (tu równa 4), zaś głębokość drzewa to liczba krawędzi między wierzchołkiem a najbardziej odległym liściem (tu równa 3).



Rys. 1. Przykład drzewa klasyfikacyjnego; źródło: opracowanie własne

Krótką charakterystyka klasycznych metod drzewkowych

Zdecydowana większość algorytmów podziału drzew klasyfikacyjnych i regresyjnych wywodzi się z 3 klasycznych metod:

- ◆ CLS (*Concept Learning System*),
- ◆ AID (*Automatic Interaction Detection*),
- ◆ CART (*Classification and Regression Trees*).

Rodzina CLS

Algorytm CLS powstał w latach 60. ubiegłego stulecia i służył do binarnego podziału zbioru obiektów na dwie klasy: klasę pozytywną (K^+) i klasę negatywną (K^-). CLS wybierał tę cechę, która najlepiej różnicowała obiekty należące do K^+ i K^- . Kolejne algorytmy oparte na metodzie CLS to m.in.: ID3 (*Iterative Dichotomizer*), ASSISTANT i C4.5. Ten ostatni, chociaż popularny wśród naukowców, nie zyskał sympatii specjalistów ds. marketingu. Dzieje się tak zapewne ze względu na fakt, że jego implementacja pozwala na tworzenie zestawu reguł, a pomija graficzny wynik analizy, jakim jest drzewo.

Rodzina AID

AID to metoda detekcji interakcji, która pojawiła się na początku lat 60., jako alternatywny dla regresji sposób analizowania danych. Z czasem okazało się, że metoda ta nie jest



pozbawiona pewnych wad, co spowodowało ewolucję algorytmu w następnych dziesięcioleciach. Część kolejnych modyfikacji (MAID, XAID) służyła do budowy modeli regresyjnych, a część (THAID, CHAID) do budowy modeli dyskryminacyjnych. Najnowszą odmianą metody jest opracowana w 2000 r. WAID.

Obecny w pakiecie *STATISTICA Data Miner* algorytm CHAID wykorzystuje test niezależności chi-kwadrat i mnożnik Bonferroniego. Na każdym etapie podziału drzewa tworzy się tabelę kontyngencji, w której zestawia się zmienną zależną i predyktor. Jeśli zmienna zależna ma $d \geq 2$ kategorii, a predyktor $c \geq 2$ kategorii, to dąży się do redukcji tabeli kontyngencji o wymiarach $d \times c$ do bardziej istotnej¹ o wymiarach $d \times j$, przez łączenie w dozwolony sposób kategorii predyktora. Oryginalny CHAID pozwala budować modele dyskryminacyjne, czyli takie, których zmienna zależna jest zmienną nominalną.

Inna modyfikacja – XAID (Exhaustive CHAID) – pozwala na tworzenie drzew klasyfikacyjnych i regresyjnych. Jeśli chodzi o te pierwsze, to wykorzystano algorytm Kassa – CHAID, jeśli o drugie, to statystykę F. Wprowadzono też inne usprawnienie – metodę *Monte Carlo*. Zdaniem autorów ma ona na celu szacowanie stabilności modelu dla małych i dużych prób, a ponadto zapewnia „sprawiedliwy” dobór predyktorów (bez dyskryminacji zmiennych wielowariantowych).

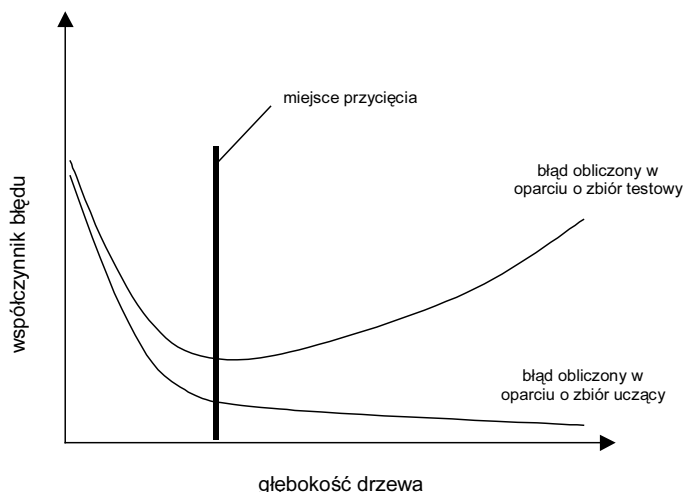
Cechami charakterystycznymi metod CHAID i XAID są:

- ♦ budowa modeli dyskryminacyjnych i regresyjnych,
- ♦ możliwość podziału węzła na więcej niż dwa podzbiory,
- ♦ brak możliwości przycinania drzewa – w przeciwieństwie do CART, algorytm przestaje dzielić zbiór obserwacji, gdy kryterium stopu nie pozwala na dalszy „rozwrost” drzewa.

CART

CART to najbardziej zaawansowana metoda budowy drzew klasyfikacyjnych i regresyjnych. Powstała na początku lat 80. ubiegłego stulecia i do dziś doczekała się kilku nieznacznych modyfikacji. Reguły podziału drzewa używane w tej metodzie to: indeks Giniego, miara entropii i reguła podziału na dwie części (*twoing rule*). Cechą charakterystyczną metody jest nadmierny rozrost drzewa i przycinanie (*pruning*) poszczególnych gałęzi w celu redukcji opisu liści (przy nieznacznym wzroście błędu klasyfikacji). Pozwala to na porównanie modelu rozbudowanego i modelu ze zredukowaną liczbą węzłów, czasami bowiem o jakości drzewa nie decyduje trafność predykcji, ale przydatność wygenerowanych reguł. Sprawdza się, jaka jest różnica między błędem klasyfikacji całego drzewa (α_0) a błędem klasyfikacji drzewa z usuniętą gałęzią (α_1). Po pierwszym etapie przycinania wybiera się to α_1 , dla którego różnica ta jest najmniejsza. W kolejnym kroku drzewem wyjściowym jest już α_1 i to jego błąd jest porównywany z błędem dla drzewa α_2 . Wybiera się to α_2 , dla którego różnica w błędach klasyfikacji między α_2 a α_1 jest najmniejsza.

¹ Z punktu widzenia testu niezależności chi-kwadrat



Rys. 2. Przykład przycinania z uwzględnieniem różnych współczynników błędu; opracowanie własne

Drugą ważną zaletą algorytmu jest jednocześnie zestawienie kosztu resubstytucji (współczynnika błędu obliczonego ze zbioru uczącego) ze współczynnikiem błędu obliczonym na zbiorze testowym (ta druga wartość może być wynikiem prostej walidacji, walidacji krzyżowej, wielokrotnej walidacji krzyżowej czy metod bootstrapowych). Przycinanie drzewa dokonywane jest z uwzględnieniem obu współczynników (rys. 2). Wymienione wyżej cechy algorytmu powodują, że jest najchętniej stosowanym „drzewkowym” narzędziem *data mining*.

Predykcja lojalności klientów za pomocą metody CART

Decydując się na analizę danych za pomocą metod drzewkowych, trzeba uwzględnić następujące etapy:

- ◆ wybór reguły podziału,
- ◆ ustalenie prawdopodobieństwa a priori występowania klas,
- ◆ wybór kryterium stopu,
- ◆ zaawansowane szacowanie błędu klasyfikacji z wykorzystaniem zbioru uczącego i testowego,
- ◆ interpretacja liści, czyli zamiana modelu drzewkowego na zestaw reguł typu „jeśli zdanie Z_1 , to zdanie Z_2 ”.

W analizie wykorzystano zbiór obserwacji z rozdziału poświęconego analizie rzetelności skal satysfakcji i lojalności. Analiza głównych składowych pozwoliła pogrupować 16 stwierdzeń ze skali SERVQUAL w 4 czynniki. Oprócz zmiennych demograficznych (płci, wykształcenia, wieku i miejsca zamieszkania) włączono do obliczeń stwierdzenie –

reprezentanta każdego czynnika. O tym, którą zmienną wybrać, decydowała wartość ładunku czynnikowego (wybierano najwyższą), dlatego też wybrano:

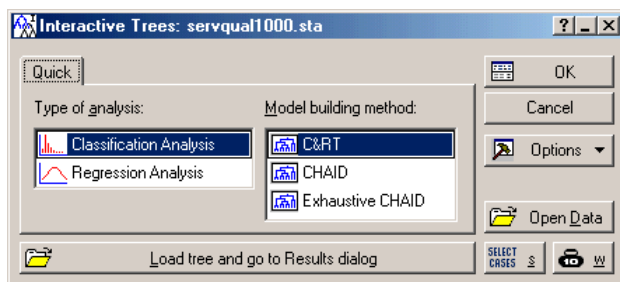
- ♦ z czynnika 1. zmienną K1 (wart. ładunku czynnikowego = 0,820711): „Firma X posiada nowoczesnie wyglądający sprzęt”,
- ♦ z czynnika 2. zmienną R2 (wart. ładunku czynnikowego = 0,813690): „Firma X już za pierwszym razem wykonuje usługę właściwie”,
- ♦ z czynnika 3. zmienną L2 (wart. ładunku czynnikowego = 0,791467): „Będę polecać firmę X moim znajomym”,
- ♦ z czynnika 4. zmienną E1 (wart. ładunku czynnikowego = 0,814275): „Firma X traktuje Pana(Panią) indywidualnie”.

Zmienną zależną będzie reprezentant czynnika określającego stopień lojalności respondenta – stwierdzenie L2, zaś predyktorami pozostałe 7 zmiennych: 4 z nich to zmienne kategoryjne (cechy demograficzne respondenta), a 3 to zmienne przedziałowe (wybrane stwierdzenia ze skali SERVQUAL).

Stwierdzenie L2 dotyczy polecenia firmy znajomym. Skala jest 7-punktowa, gdzie 1 to wariant „zdecydowanie się nie zgadzam”, a 7 to wariant „zdecydowanie się zgadzam”. Zredukowano poziom pomiaru zmiennej z przedziałowego (1–7) na nominalny. Połączono oceny 1 i 2 ze skali SERVQUAL w jedną kategorię – osób „zdecydowanie nie zamierzających polecać usług firmy”, oceny 3, 4 i 5 połączono w drugą kategorię – osób „niezdecydowanych”, natomiast oceny 6 i 7 połączono w trzecią kategorię – osób „zdecydowanie zamierzających polecić usługi firmy znajomym”.

System *STATISTICA Data Miner* zawiera 4 moduły do analizy drzewkowej:

- ♦ ogólne modele drzew klasyfikacyjnych i regresyjnych (moduł oparty o algorytm CART),
- ♦ ogólne modele CHAID,
- ♦ drzewa interakcyjne (C&RT, CHAID i XAID),
- ♦ drzewa klasyfikacyjne i regresyjne ze wzmocnieniem (ang. boosted).



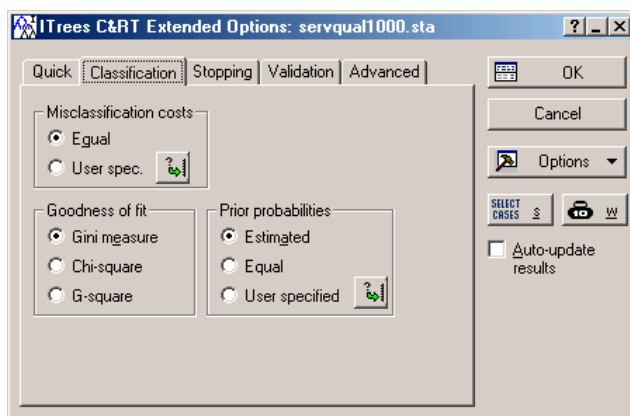
Rys. 3. Panel początkowy modułu „Drzewa interakcyjne (C&RT, CHAID)” z pakietu *STATISTICA*

W analizie wykorzystano algorytm CART z modułu drzew interakcyjnych (rys. 3), jako że pozwala on na ręczne sterowanie wielkością i głębokością drzewa. Ten pierwszy parametr

można zmieniać, usuwając poszczególne krawędzie (gałęzie) modelu, zaś ten drugi – usuwając cały etap podziału jednocześnie.

Przed przystąpieniem do analizy wybrano następujące ustawienia (rys. 4):

- ♦ reguła podziału – miara Giniego,
- ♦ prawdopodobieństwa a priori – szacowane z próby uczącej,
- ♦ koszty błędnej klasyfikacji – równe,
- ♦ kryterium stopu: przy błędnej klasyfikacji,
- ♦ minimalna liczebność węzła końcowego $n=30$,
- ♦ maksymalna liczba poziomów (głębokość) drzewa $n=10$,
- ♦ maksymalna liczba węzłów $n=1000$,
- ♦ szacowanie błędu za pomocą 10-krotnej walidacji testowej.



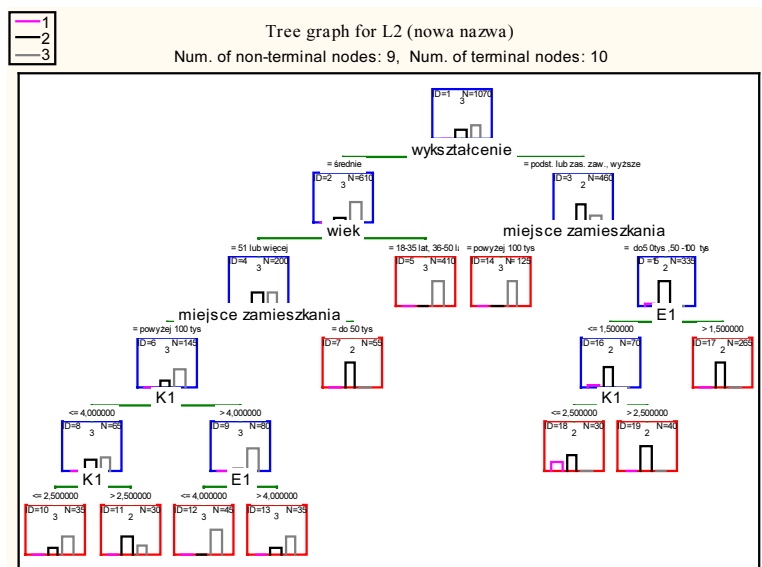
Rys. 4. Ustawienia początkowe analizy CART w module „Drzewa interakcyjne (C&RT, CHAID)”

Graficzny wynik analizy to drzewo z rys. 5., zawierające 10 liści i głębokie na 5 poziomów. Przy budowie modelu wykorzystano 5 zmiennych niezależnych: wykształcenie, miejsce zamieszkania, wiek, ocenę indywidualnego traktowania klienta (E1) i ocenę nowoczesności sprzętu (K1).

Jak łatwo zauważyć, węzeł nr 9 dzieli się na węzły 12. i 13., które zawierają klasę 3., czyli respondentów zdecydowanych polecić usługi firmy znajomym. Podobnie z węzłem nr 16, z którego wydzielono 2 liście (18. i 19.) zawierające tę samą klasę zmiennej zależnej oraz z węzłem nr 15 – również podzielonym na dwie takie same klasy (16. i 17.).

Powstaje pytanie, dlaczego dokonano podziału, skoro wydzielone liście zawierają tę samą klasę? Odpowiedź jest prosta. Po pierwsze, celem podziału jest wydzielenie maksymalnie homogenicznych liści z punktu widzenia wartości zmiennej zależnej. Węzeł 9. zawiera 10 przypadków klientów „niezdecydowanych” i 70 „zdecydowanych polecić usługi znajomym”, natomiast węzeł 12. zawiera tylko klientów zdecydowanych (45 przypadków),

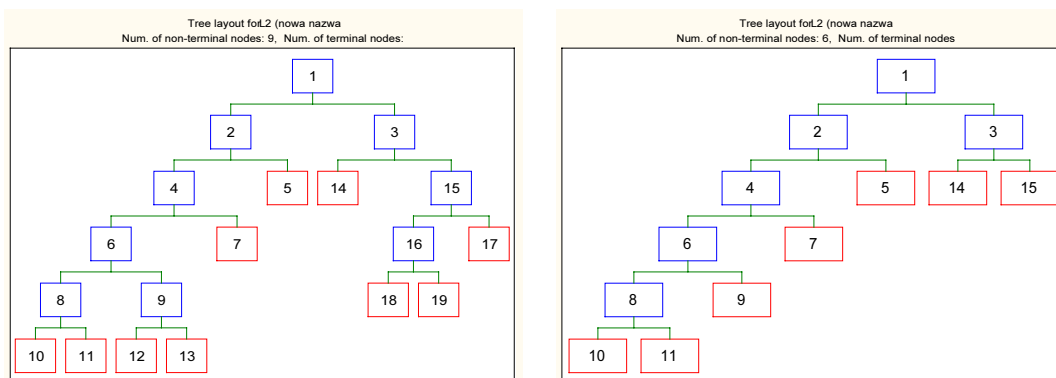
a węzeł 13. – 10 „niezdecydowanych” i 25 „zdecydowanych”. O dalszym podziale zdecydowała tutaj 100-procentowa jednorodność liścia nr 12.



Rys. 5. Drzewo klasyfikacyjne CART

Po drugie, w nazewnictwie liści stosuje się zasadę majoryzacji zbiorów, polegającą na przyjęciu nazwy od klasy najliczniej występującej w liściu. O ile liść nr 12 zawiera wyłącznie respondentów z klasy 3. (liść przejmuje nazwę tej klasy), o tyle w liściu nr 13 jest 10 przypadków klasy 2. i 25 przypadków z klasy 3. Gdyby nie zasada majoryzacji, to liść ten można by nazwać „w 71% klasa 3”.

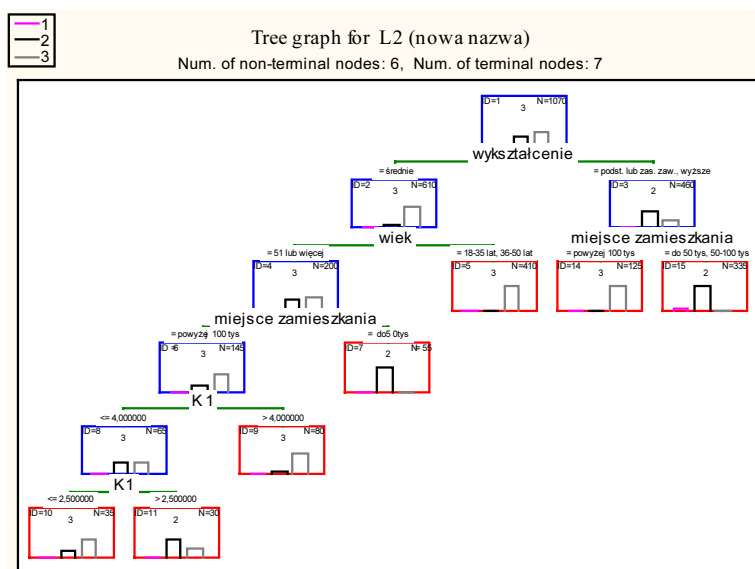
W zależności od celu badań można przyjąć, że te podziały są zbędne. Moduł *Drzewa interakcyjne* (C&RT, CHAID) pozwala zredukować model o dowolne gałęzie lub poziomy bez ingerencji w pozostałą strukturę drzewa. Zmianę układu drzewa przedstawiono na rys. 6.



Rys. 6. Układ pierwotnego drzewa klasyfikacyjnego CART (z lewej) i drzewa z usuniętymi gałęziami w węzle 9. i 15. (z prawej)

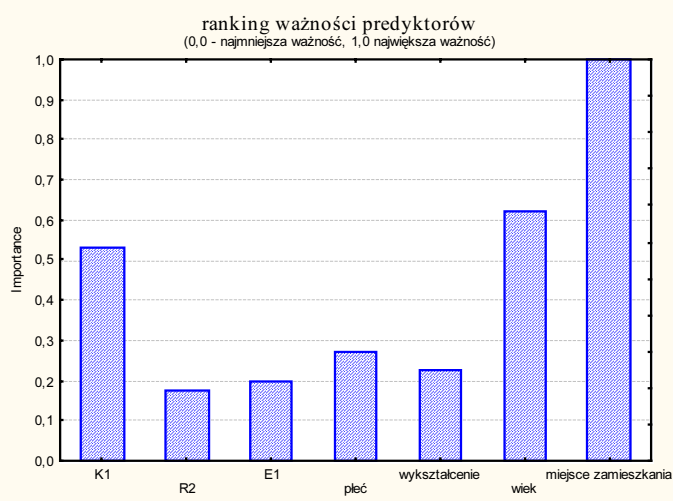


Zredukowane drzewo (rys. 7.) ma teraz 7 liści i jest głębokie na 5 poziomów.



Rys. 7. Drzewo klasyfikacyjne CART ze zredukowaną liczbą gałęzi

Współczynnik błędu po 10-krotnej walidacji krzyżowej wynosi 0,050916, co oznacza, że model jest bardzo dobrze dopasowany do danych. Ranking ważności predyktorów (rys. 8) przedstawia zmienne niezależne, które porangowano na skali od 0,0 do 1,0, gdzie wartości najwyższe oznaczają największy wpływ danej zmiennej na zmienną zależną. W tym przypadku o poleceniu usług firmy znajomym decydują przede wszystkim: miejsce zamieszkania, wiek i ocena nowoczesności sprzętu dokonana przez respondenta.



Rys. 8. Ranking ważności predyktorów dla zredukowanego drzewa klasyfikacyjnego CART



Liczba liści przekłada się bezpośrednio na liczbę reguł typu „jeśli zdanie Z_1 , to zdanie Z_2 ”. Ten spójnik międzyzdanowy jest odpowiednikiem funktora prawdziwościowego implikacji „ \supset ” i oznacza zazwyczaj zdanie warunkowe. Używając tego spójnika w mowie potocznej, przyjmuje się, że między zdaniami składowymi istnieje powiązanie rzeczowe lub formalne, tzn. pierwsze zdanie Z_1 implikuje drugie Z_2 . W przypadku reguł indukcyjnych opartych na drzewach klasyfikacyjnych związek między poprzednikiem Z_1 (poprzednikami) a następnikiem Z_2 może mieć charakter przyczynowy lub strukturalny (taki, który wynika z rozmieszczenia przedmiotów w przestrzeni albo zdarzeń w czasie). Jeśli zdania opisujące reguły indukcyjne składają się z kilku poprzedników, to zawierają również słowny odpowiednik funktora prawdziwościowego koniunkcji „ \wedge ”. Poprzedniki te łączy się spójnikiem międzyzdanowym „i”, który ma tutaj znaczenie syntetyzujące, np. „*jeżeli Z_1 i Z_2 i Z_3 , to Z_4* ”. Oznacza to, że jeżeli spełnione zostały warunki Z_1 i Z_2 i Z_3 łącznie, to wystąpi Z_4 . Gdyby zastosować spójnik międzyzdanowy „oraz”, to zdanie to należałoby rozumieć następująco: jeśli występuje Z_1 , to występuje Z_4 , oraz jeśli występuje Z_2 , to występuje Z_4 , oraz jeśli występuje Z_3 , to występuje Z_4 . Spójnik ten ma znaczenie enumeracyjne, tzn. poprzedniki nie są traktowane łącznie.

Przykładowe reguły brzmią następująco:

1. Jeśli respondent ma wykształcenie średnie i co najwyżej 50 lat, to zdecydowanie poleci usługi firmy swoim znajomym (węzeł nr 3),
2. Jeśli respondent ma wykształcenie średnie i ma co najmniej 51 lat i mieszka w miejscowości o liczbie mieszkańców do 50 tys., to nie jest pewien, czy poleci usługi firmy swoim znajomym (węzeł nr 7),
3. Jeśli respondent ma wykształcenie średnie i ma co najmniej 51 lat, i mieszka w miejscowości o liczbie mieszkańców powyżej 100 tys., i ocenia sprzęt w firmie jako nowoczesny (ocena > 4,0), to zdecydowanie poleci usługi firmy swoim znajomym (węzeł nr 9),
4. Jeśli respondent ma wykształcenie inne niż średnie i mieszka w miejscowości o liczbie mieszkańców do 100 tys., to nie jest pewien czy poleci usługi firmy swoim znajomym (węzeł nr 15).

Gdyby badacz uznał, że część z tych reguł jest z praktycznego punktu widzenia nieprzydatna, może je odrzucić przez redukcję kolejnych gałęzi i poziomów modelu. W module *Drzewa interakcyjne (C&RT, CHAID)* jest to sprawą stosunkowo prostą.

Podsumowanie

W badaniach marketingowych bardzo często analizuje się zmienne wyrażone na skalach słabych. Mogą to być zarówno zmienne binarne (np. kupi – nie kupi), jak też zmienne wielokategorialne (np. kto kupi? na jaką cechę produktu zwróci uwagę? jaką markę wybierze?). Poziom pomiaru zmiennych zależnych ogranicza możliwość stosowania wielu narzędzi statystycznych – niezajdujących zastosowania w predykcji jakościowej.



Elastycznym i łatwym w użyciu narzędziem do analizy takich zbiorów obserwacji są drzewa klasyfikacyjne i regresyjne. Elastyczność dotyczy poziomów pomiaru zmiennych zależnych i zmiennych niezależnych. Łatwość w użyciu związana jest z przejrzystością graficznego wyniku analizy – drzewa oraz z brakiem ograniczeń dotyczących rozkładów zmiennych.

Do innych ważnych zalet zaprezentowanego w niniejszej pracy algorytmu CART należy dodać:

- ♦ niewrażliwość na występowanie obserwacji nietypowych, co do których istnieje przypuszczenie, że pochodzą z innej populacji; klasyczne statystyczne metody analizy nie są odporne na występowanie takich przypadków; zaletą tej metody jest szczególnie ważna, jeśli wykorzystuje się regresyjne właściwości CART;
- ♦ możliwość skutecznego wykorzystania w zbiorach danych cechujących się licznymi brakami danych w zmiennych niezależnych;
- ♦ wykorzystywanie tych samych zmiennych w różnych częściach drzewa (pozwala odkryć kontekst zależności i interakcji między zmiennymi);
- ♦ możliwość wykorzystania liniowej kombinacji zmiennych ilościowych w celu określenia dalszego podziału drzewa.

Literatura

1. Breiman L., Friedman J.H., Olshen R.A., Stone C.J.; *Classification and Regression Trees*; Chapman and Hall; 1993.
2. Gatnar E., Nieparametryczna metoda dyskryminacji i regresji, PWN, Warszawa 2001.
3. Łapczyński M., *Detekcja interakcji w modelach drzewkowych – próba syntezy*, referat z VII Warsztatów Metodologicznych pt. „Badanie czynników wpływających na zachowania podmiotów rynkowych. Analiza interakcji”, Wrocław 15 maja 2003.
4. Łapczyński M., *Przyczynowa interpretacja drzew klasyfikacyjnych*, [w:] *Zależności przyczynowo-skutkowe w badaniach rynkowych i marketingowych* pod red. S. Mynarskiego, Wydawnictwo AE w Krakowie, Kraków 2002, s. 47-60.