Studia Informatica Pomerania nr 1/2016 (39) | www.wnus.edu.pl/si | DOI: 10.18276/si.2016.39-05 | 49–59

ALGORYTMY DO KONSTRUOWANIA DRZEW DECYZYJNYCH W PRZEWIDYWANIU SKUTECZNOŚCI KAMPANII TELEMARKETINGOWEJ BANKU

JAN KOZAK[1], PRZEMYSŁAW JUSZCZUK[2]

[1] Uniwersytet Ekonomiczny w Katowicach
Wydział Informatyki i Komunikacji
e-mail: jan.kozak@ue.katowice.pl
[2] Uniwersytet Śląski
Wydział Informatyki i Nauki o Materiałach
e-mail: przemyslaw.juszczuk@us.edu.pl

SŁOWA KLUCZOWE

drzewa decyzyjne, analiza danych, kampania telemarketingowa

STRESZCZENIE

W artykule dokonano analizy kampanii telemarketingowej portugalskiego banku. Dane zawierają 17 atrybutów (cech), w tym informację o skuteczności przeprowadzonej rozmowy związanej z ofertą lokaty bankowej. Analiza przeprowadzona została z zastosowaniem algorytmów do konstruowania drzew decyzyjnych (m.in. CART, C4.5), a w jej wyniku, na podstawie wartości cech klienta, wykonana została predykcja określająca skutek rozmowy telemarketingowej. Wykonane doświadczenia pozwoliły na analizę wyników poszczególnych klasyfikatorów pod względem różnych miar oceny jakości klasyfikacji. Jest to szczególnie ważne w przypadku rzeczywistych zbiorów danych z nierównomiernie rozłożonymi klasami decyzyjnymi.

Wprowadzenie

W dzisiejszych czasach banki mają kilka możliwości przeprowadzania kampanii marketingowych swoich produktów. Do najważniejszych należy zaliczyć kampanie masowe oraz kampanie skierowane (bezpośrednie). Kampanie masowe skierowane są do całej grupy osób i obecnie przynoszą bardzo małe korzyści – mniej niż 1% pozytywnego odzewu (Moro, Laureano, Cortez, 2011). W związku z tym banki coraz częściej wykorzystują możliwość stosowania kampa-

nii skierowanych (bezpośrednich), polegających na bezpośrednim proponowaniu konkretnemu klientowi skorzystania z danego produktu. Duże zbiory danych związane z klientami, a także informacje o wcześniejszych działaniach, pozwalają w coraz większym stopniu stosować w bankach metody eksploracji danych (ang. *data mining*) w celu dopasowania oferty marketingowej do profilu klientów.

Stosowanie ukierunkowanych kampanii marketingowych jest szczególnie ważne w przypadku kryzysu, z jakim obecnie mierzy się rynek bankowości, nieufności klientów i dużej konkurencyjności. Należy pamiętać, że błędne skierowanie oferty do klienta może zniechęcić go do przyszłej oferty banku. Dlatego niezwykle ważne jest zastosowanie metod eksploracji danych do predykcji odpowiedzi klienta i wcześniejsze wykluczenie skierowania kampanii do klientów, którzy potencjalnie nie skorzystają z oferty.

W artykule przeprowadzona zostanie analiza kampanii telemarketingowej portugalskiego banku w celu wyznaczenia klasyfikatora pozwalającego na dobrą predykcję skuteczności kampanii. Analiza podobnych zbiorów danych jest trudna, ponieważ zazwyczaj posiadają one dużą liczbę atrybutów (w tym ciągłych) i nierówno rozłożone przypadki w klasach decyzyjnych. Również w omawianej sytuacji ponad 88% przypadków w zbiorze danych zakończyła się nieskorzystaniem przez klienta z oferty, a tylko niespełna w 12% kampania zakończyła się pozytywnie. Wobec tego należy stosować wybrane miary oceny jakości klasyfikacji niwelujące sytuację, w której wszystkie kampanie zostaną z góry ocenione jako bezzasadne (Boryczka, Kozak, 2014).

Zbiór danych – kampania telemarketingowa banku

W artykule analizie poddano zbiór danych dotyczący skierowanej (bezpośredniej) kampanii marketingowej portugalskiego banku. Kampanie te polegały na przeprowadzaniu rozmów telefonicznych z klientami banku w celu zaoferowania lokaty terminowej, w związku z czym wymagały również wielokrotnych połączeń z tym samym klientem. Dane zostały opracowane przez S. Moro, R. Laureano i P. Corteza i opisane w artykule (Moro i in., 2011). Następnie podlegały wielokrotnej analizie zarówno pod względem predykcji skuteczności kampanii, jak i stosowania innych metod eksploracji danych (niezwiązanych bezpośrednio z klasyfikacją) (Elsalamony, 2014). Ostatecznie w 2014 roku zaproponowany został nowy zestaw danych (Moro, Cortez, Rita, 2014).

W artykule analizie poddano 45 211 przypadków opisanych 17 atrybutami (w tym atrybut decyzyjny) bez brakujących atrybutów. Dane zapisane są w dwóch zbiorach danych przygotowanych przez autorów artykułu (Moro i in., 2011) i w celach porównawczych, na potrzeby tego artykułu nie dokonywano żadnych modyfikacji.

Dane zawierają informacje:

a) o klientach banku (8 cech):

– wiek klienta (atrybut numeryczny, bez dyskretyzacji),
– praca (atrybut kategoryczny – 12 wartości),
– stan cywilny (atrybut kategoryczny – 3 wartości),

– wykształcenie (atrybut kategoryczny – 4 wartości),
– informacje o zaległościach kredytowych (atrybut binarny – tak/nie),
– średnie roczne saldo (atrybut, bez dyskretyzacji),
– informacje o kredycie mieszkaniowym (atrybut binarny – tak/nie),
– informacje o kredycie gotówkowym (atrybut binarny – tak/nie);

b) o ostatnim kontakcie z klientem w sprawie tej kampanii (4 cechy):
– forma kontaktu (atrybut kategoryczny – 3 wartości),
– dzień (miesiąca) kontaktu (atrybut numeryczny, bez dyskretyzacji),
– miesiąc kontaktu (atrybut kategoryczny – 12 wartości),
– czas trwania kontaktu (atrybut numeryczny, bez dyskretyzacji);

c) o pozostałych cechach (4 cechy):
– liczba kontaktów z klientem w sprawie tej kampanii (atrybut numeryczny, bez dyskretyzacji),
– liczba dni od ostatniego kontaktu z klientem (atrybut numeryczny, bez dyskretyzacji),
– liczba kontaktów z klientem w sprawie innych kampanii (atrybut numeryczny, bez dyskretyzacji),
– wynik poprzedniej kampanii (atrybut kategoryczny – 4 wartości);

d) oraz atrybut decyzyjny – czy klient skorzystał z produktu oferowanego w tej kampanii (atrybut binarny w przybliżonym podziale 12% – „Yes", 88% – „No").

Drzewa decyzyjne

Drzewo decyzyjne (ang. *decision tree*) to acykliczny graf skierowany, w którym wierzchołki nazywane są węzłami, krawędzie gałęziami, wierzchołki nieposiadające potomków liśćmi, a wierzchołek nieposiadający rodzica korzeniem. Wszystkie węzły zawierają testy na atrybutach warunkowych powstałe zgodnie z przyjętym kryterium podziału. Testy te dokonują podziału danych w zależności od wartości ich atrybutów (cech), a każdy wynik testu reprezentowany jest przez gałęzie.

Prosta i intuicyjna budowa drzewa decyzyjnego sprawia, że drzewa o niedużej wielkości mogą być analizowane bezpośrednio przez użytkownika, natomiast w przypadku zastosowania drzewa o dużej wielkości klasyfikacja i tak jest znacznie szybsza niż przy innych metodach. Ponadto drzewa decyzyjne można stosunkowo łatwo zapisać w postaci reguł decyzyjnych, co pozwala na korzystanie z ich rezultatów również w systemach ściśle związanych z regułami decyzyjnymi. Ta przewaga nad innymi klasyfikatorami motywuje do dalszych prac nad udoskonalaniem algorytmów do konstruowania drzew decyzyjnych, w celu wyeliminowania wszystkich niedogodności i poprawy jakości budowanych drzew decyzyjnych. Dodatkowo zastosowanie rodziny klasyfikatorów w postaci lasów drzew decyzyjnych wydaje się szczególnie przydane podczas budowania drzew decyzyjnych metodami stochastycznymi (specjalnie opracowanymi heurystykami).

Konstruowanie drzewa decyzyjnego oparte jest na zasadzie „dziel i zwyciężaj" i polega na wielokrotnym, rekursywnym podziale danych, co powoduje rozdzielenie problemu na mniejsze „podproblemy". Standardowo podział odbywa się zachłannie, a więc wybierany jest potencjalnie najlepszy podział pod względem wartości wyznaczonych na podstawie wybranego kryterium podziału. Dobre kryterium podziału powinno jak najmniej różnicować obiekty (pod względem ich klasy decyzyjnej) w każdym potomku węzła. W momencie kiedy w węźle wszystkie obiekty należą do jednej klasy decyzyjnej, węzeł ten staje się etykietą klasy (liściem). W efekcie drzewo decyzyjne reprezentuje proces podziału obiektów (pod względem wartości ich cech) ze zbioru danych na jednorodne klasy. Reguła podziału powinna minimalizować błąd klasyfikacji przypadków ze zbioru testowego.

Drzewa decyzyjne konstruowane są według tzw. metody zstępującej (ang. *top-down*), czyli pierwszy (potencjalnie najlepszy) atrybut i wartości, według których podzielone zostaną dane stanowią korzeń drzewa, węzły potomne dokonują kolejnego podziału według tej samej zasady. Dzielą w ten sposób dane treningowe na kolejne części, schodząc w dół drzewa, aż do osiągnięcia kryterium stopu i ustalenia wartości atrybutu decyzyjnego w liściu drzewa. Przeważnie po etapie konstruowania drzewa decyzyjnego wykonywane jest tzw. przycinanie, którego celem jest przede wszystkim zapobieganie przetrenowaniu budowanego klasyfikatora. Przykładem przycinania drzew decyzyjnych jest metoda wstępująca (ang. *bottom-up*), gdzie kolejne węzły sprawdzane od dołu drzewa porównywane są pod względem dokładności klasyfikacji z liściem aktualnej ścieżki w drzewie. Jeśli szacunkowy błąd klasyfikacji węzła w stosunku do liścia mieści się w określonym przedziale, to węzeł drzewa zostaje zastąpiony liściem.

Kryterium podziału

Kryterium podziału (ang. *splitting rule*) stosowane jest w celu znalezienia najlepszego testu, który podzieli zbiór danych w węźle na dwa (lub więcej, w zależności od typu drzewa decyzyjnego) podzbiory danych. Testem określany jest warunek dla podziału danych. Warunek ten jest ściśle związany z atrybutami oraz wszystkimi możliwymi wartościami tych atrybutów.

Wybór podziału danych w każdym węźle jest zdecydowanie najtrudniejszym i najbardziej złożonym etapem konstruowania drzew decyzyjnych. Zastosowanie konkretnego kryterium zależne jest od stosowanego algorytmu lub nawet konkretnych zastosowań algorytmu.

Przykładowo, w przypadku algorytmu CART w celu oceny testu przeważnie wyznaczana jest miara nieczystości $i(t)$ (ang. *impurity function*), która określa maksymalną jednorodność węzłów potomnych. Ponieważ miara nieczystości węzła nadrzędnego m_p jest stała dla każdego z możliwych podziałów $a_j \leq a_j^R$, j = 1, …, M (gdzie M oznacza liczbę atrybutów, a a_j^R to najlepszy podział dla atrybutu a_j), maksymalna jednorodność lewego i prawego potomka będzie określona przez maksymalną różnicę miary nieczystości $\Delta i(t)$ (Timofeev, 2004):

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r) \qquad (1)$$

gdzie:

P_l – prawdopodobieństwo przejścia obiektu do węzła m_l (lewego poddrzewa),

P_r – prawdopodobieństwo przejścia obiektu do węzła m_r (prawego poddrzewa).

W związku z tym algorytm do konstruowania drzewa decyzyjnego, przy wyborze podziału dla każdego węzła, rozwiązuje problem maksymalizacyjny. Polega on na przeszukaniu wszystkich możliwych wartości atrybutów w celu znalezienia najlepszego podziału (największej wartości miary różnorodności, a co za tym idzie różnicy miary nieczystości) (Timofeev, 2004):

$$[\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r)]. \qquad (2)$$

Dla algorytmu CART Breiman, Friedman, Olshen, Stone (1984) zaproponowali dwa kryteria podziału, czyli sposoby wyznaczania miary różnorodności: Giniego oraz podziału na dwie części. Obydwie przedstawione poniżej reguły zawarte zostały (osobno) w funkcji heurystycznej proponowanego algorytmu. Inne kryteria miary różnorodności oparte są m.in. na entropii (stosowane w algorytmie C4.5), proporcji błędnych klasyfikacji, rozkładzie chi-kwadrat i wielu innych podejściach dokładniej opisanych w książkach (Koronacki, Ćwik, 2008; Rokach, Maimon, 2008).

Kryterium podziału Giniego (ang. *Gini splitting rule*) oparte została na indeksie Giniego, czyli mierze koncentracji zmiennej losowej. Nadrzędnym celem w tym przypadku jest dokonanie podziału na możliwie jednorodne przypadki w węzłach potomnych. Miara nieczystości wyznaczana jest na podstawie wzoru:

$$i(t) = \sum_{k \neq o} p\,(oVm)p(kVm) \qquad (3)$$

gdzie:

$p(k|m)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej k w węźle m,

$p(o|m)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej o w węźle m,

o i k – klasy decyzyjne.

Warunek, według którego dokonywany jest podział, wyznaczany na podstawie wzorów (1) i (2), tworzy następującą formułę (Breiman i in., 1984; Timofeev, 2004):

$$(-\sum_k^K = {}_1p^2\,(kVm_p) + P_l \sum_k^K = {}_1p^2\,(kVm_l) + P_r \sum_k^K = {}_1p^2\,(kVm_r)) \qquad (4)$$

gdzie:

$p(kVm_p)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej k w węźle m_p,

$p(kVm_l)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej k w węźle m_l,

$p(kVm_r)$ – prawdopodobieństwo wystąpienia klasy decyzyjnej k w węźle m_r,

K – liczba klas decyzyjnych.

Kryterium podziału na dwie części (ang. *twoing rule*) przede wszystkim dokonuje podziału danych na dwie możliwie równe części (dwa podzbiory). Jednorodność klasy decyzyjnej jest w tym przypadku mniej znacząca niż podczas stosowania kryterium Giniego, choć odgrywa pewną rolę. Miara różnorodności jest tu określona jako:

$$\Delta i(t) = \frac{P_l P_r}{4}\,[\sum_k^K |p(k|m_l) - k|m_r)|]^2 \qquad (5)$$

Warunek, według którego dokonywany jest podział wyznaczany na podstawie wzorów (1) i (2), można zapisać jako (Breiman i in., 1984; Timofeev, 2004):

$$(\frac{P_l P_r}{4}[\sum^K{}_{k=l} |p(k|m_l) - k|m_r)|]^2 \qquad (6)$$

Często stosowanymi kryteriami podziału są również reguły oparte na entropii, jak np. znany z algorytmu ID3 (Quinlan, 1986) zysk (przyrost) informacji (ang. *information gain*)

lub w szczególności (algorytm C4.5) reguła względnego zysku (ang. *gain ratio*), zwana także współczynnikiem przyrostu informacji i stosowana w algorytmie C4.5 (Quinlan, 1993). W przypadku zastosowania tych kryteriów budowane drzewa decyzyjne niekoniecznie są drzewami binarnymi, ponieważ testy w węzłach odpowiadają atrybutom, a gałęzie możliwym wartościom tych atrybutów (dla danych dyskretnych). Dla każdego węzła wybierany jest podział o najwyższej wartości względnego zysku informacji:

$$\left(\frac{\text{zyskInf}(a_i,\text{S})}{\text{entropa }(a_i,\text{S})}\right) \qquad (7)$$

gdzie zyskInf(a_i, S) jest zyskiem informacji (8), a entropa (a_i, S) jest entropią rozkładu danych ze zbioru S na podstawie wartości atrybutu a_i (wzór (9)).

$$\text{zyskInf}(a_i, \text{S}) = \text{entropia}(\text{y,S}) - \sum_{k=1}^{K} \frac{|S_k|}{|S|} \text{entropia}(\text{y}, \text{S}_k) \qquad (8)$$

$$\text{entropia}(\text{y}, \text{S}) = \sum_{j=1}^{|y|} \frac{-|S_j|}{|S|} \cdot \log_2 \frac{|S_j|}{|S|} \qquad (9)$$

Algorytmy do konstruowania drzew decyzyjnych

W literaturze można znaleźć wiele algorytmów do konstruowania drzew decyzyjnych, z których do najpopularniejszych należą algorytmy CART oraz C4.5 (a w jego następstwie C5.0). Wyróżnić należy również takie algorytmy, jak: CHAID zaproponowany w 1980 roku algorytm (Kass, 1980), w którym dla wyznaczenia każdego podziału stosuje się niezależność chi-kwadrat oraz mnożnik Bonferroniego; QUEST zaproponowany w 1997 roku algorytm (Loh, Shih, 1997), w którym zastosowano parametryczne metody statystyczne. Wyróżnia się ponadto wiele algorytmów stosowanych do budowy drzew decyzyjnych, a ich przegląd oraz dokładne porównanie znajduje się w m.in. w książce (Lim, Loh, Shih, 2000).

Algorytm CART

Algorytm CART, zaproponowany przez Breimana i in. (Breiman i in., 1984), jest algorytmem do konstruowania drzew klasyfikacyjnych i regresyjnych służących do budowy modeli predykcyjnych i deskryptywnych. Drzewa klasyfikacyjne stosowane są wówczas, gdy zmienna zależna (klasa decyzyjna) wyrażona jest w skali nominalnej lub porządkowej. Drzewa regresyjne stosuje się natomiast wtedy, kiedy występuje (co najmniej) przedziałowy poziom pomiaru zmiennej zależnej (wartości ciągłe dla klasy decyzyjnej). Budowa modelu predykcyjnego ma na celu predykcję jakościową lub ilościową, natomiast w przypadku budowy modelu deskryptorowego dąży się do opisu i prezentacji wzorców w badanej zbiorowości (Łapczyński, 2003).

Drzewa budowane przez algorytm CART to binarne drzewa decyzyjne zbudowane według kryterium podziału Giniego (wzór (4)) lub podziału na dwie części (wzór (6)). Skonstruowane drzewa decyzyjne podlegają przycinaniu opartym na koszcie złożoności i dopuszczają zarówno atrybuty z wartościami ciągłymi, jak i dyskretnymi. Co ciekawe, zmienna celu, czyli klasa decyzyjna – wartość w liściu drzewa, może posiadać wartości ciągłe, czyli należeć do zakresu liczb

rzeczywistych. Algorytm CART konstruuje w takim przypadku tzw. drzewo regresyjne. Dane zastosowane do uczenia drzewa oraz klasyfikacji mogą posiadać brakujące wartości atrybutów.

Algorytm C4.5

Algorytm C4.5 zaproponowany przez Quinlana (Quinlan, 1996) jest udoskonaloną wersją wcześniejszego algorytmu ID3 (Quinlan, 1986). W porównaniu do algorytmu ID3 poprawione zostało m.in. kryterium podziału, tak aby uzyskiwane podziały dla większych zbiorów danych generowały mniejszy błąd klasyfikacji i możliwa była klasyfikacja obiektów z brakującymi wartościami atrybutów. W algorytmie ID3 jako kryterium podziału stosowana jest reguła zysku informacji (wzór (8)), natomiast w C4.5 reguła względnego zysku (wzór (7)).

Ponadto w algorytmie C4.5 wprowadzono przycinanie. Początkowo była to podstawowa metoda przycinania pesymistycznego (ang. *pessimistic pruning*), która następnie podlegała stopniowym udoskonaleniom (ang. *error-based pruning*). Podczas procesu uczenia się oraz klasyfikacji istnieje możliwość pracy z obiektami nieposiadającymi wartości wszystkich atrybutów (dane z brakującymi wartościami atrybutów), dodatkowo algorytm C4.5 dostosowany jest do pracy z ciągłymi wartościami atrybutów (Quinlan, 1996).

Eksperymenty

Eksperymenty wykonane zostały z zastosowaniem trzech algorytmów. Algorytmów C4.5 i CART, dokładnie opisanych w tym artykule, oraz algorytmów drzew losowych (RT – ang. *random trees*), w którym kolejne podziały wybierane są losowo – wszystkie algorytmy dostępne są w systemie WEKA (Bouckaert i in., 2013). Algorytmy przetestowano ze względu na różnego rodzaju miary jakości klasyfikatora.

Ocena jakości klasyfikacji jest jednym z problemów uczenia maszynowego. Ma zasadnicze znaczenie w kwestii stwierdzenia, czy dany klasyfikator jest dobrej, czy złej jakości. Brakuje jednak klasyfikatorów, które mogłyby być zmieniane w zależności od stosowanej miary lub też optymalizowane ze względu na kilka miar, a często w przypadku rzeczywistych problemów może okazać się, że ważniejsze są np. precyzja lub wyważenie dwóch różnych miar oceny jakości klasyfikacji (Kozak, Boryczka, 2013). W niniejszej pracy zastosowano dokładność klasyfikacji (ang. *accuracy rate*) oraz precyzję (ang. *precision*) i czułość (ang. *recall*) dla klasy „Yes", ponieważ przede wszystkim ważne jest określenie poprawnej klasyfikacji obiektów znajdującej się w mniej licznej klasie (tutaj „Yes"), a dodatkowo poprawne przewidzenie pozytywnych efektów rozmowy telemarketera.

Wszystkie te miary umożliwiają określanie jakości klasyfikacji binarnej (dla zbiorów danych z dwiema klasami decyzyjnymi). Można je wyznaczyć na podstawie macierzy błędu, umożliwiającej ocenę jakości tej klasyfikacji na podstawie informacji na temat klasy decyzyjnej obiektu oraz klasy, do której został on sklasyfikowany (Rokach, Maimon, 2008; Boryczka, Kozak, 2014).

Tabela 1. Macierz błędu – porównanie wyników algorytmów

		Predykcja „No”	Predykcja „Yes”
Algorytm C4.5			
	„No”	38 547	1 375
	„Yes”	3 112	2 177
Algorytm CART			
	„No”	38 774	1 148
	„Yes”	3 382	1 907
Algorytm drzew losowych (RT)			
	„No”	37 151	2 771
	„Yes”	2 753	2 536

Źródło: opracowanie własne.

Doświadczenia przeprowadzone zostały z zastosowaniem metody „trenuj i testuj”, gdzie jako zbiór trenujący zastosowano 4521 przypadków losowo wyselekcjonowanych przez twórców zbioru danych, natomiast jako zbiór testowy zastosowano pełen zbiór 45 211 przypadków. W tabeli 1 przedstawione zostały macierze błędu dla każdej z analizowanych metod. Na ich podstawie istnieje możliwość wyznaczenia wartości konkretnych miar klasyfikacji. W tym przypadku należy zwrócić uwagę na obciążenie zbioru danych wynikające z nierównomiernego podziału przypadków na klasy decyzyjne (ponad 88% przypadków należy do klasy „No”), dlatego w analizie poza dokładnością klasyfikacji zaproponowano precyzję i czułość. „Precyzja” dla klasy „Yes” pozwoli określić, z jaką pewnością można zakładać, że przypadek sklasyfikowany do klasy „Yes” w rzeczywistości jest w tej klasie, czyli z jakim prawdopodobieństwem można uznać, że rozmowa wskazana przez algorytm jako pozytywna w rzeczywistości zakończy się sukcesem. Natomiast „czułość” dla klasy „Yes” pozwoli określić, jak wiele przypadków należących do klasy „Yes” zostało poprawnie sklasyfikowanych, czyli jak wiele z potencjalnie pozytywnych rozmów zostało wskazanych przez algorytm. Dokładne wyniki przedstawione zostały w tabeli 2.

Tabela 2. Wyniki doświadczeń dla analizowanych algorytmów

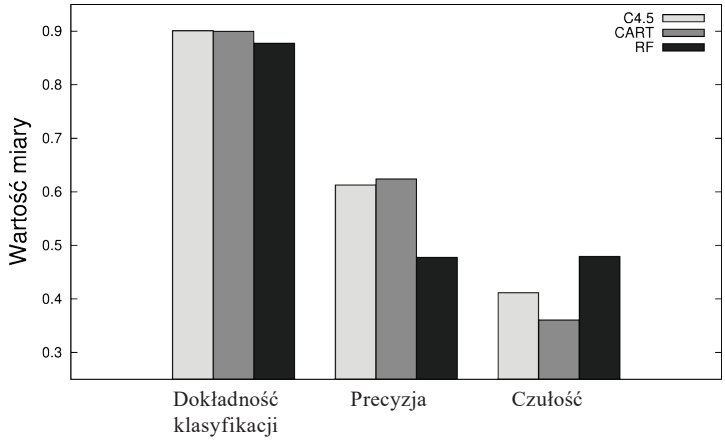
Algorytm	Dokładność klasyfikacji	Precyzja	Czułość	Liczba węzłów
C4.5	0,9008	0,6129	0,4116	146
CART	0,8998	0,6242	0,3606	19
RT	0,8778	0,4779	0,4795	1629

Źródło: opracowanie własne.

Analiza uzyskanych wyników pozwala na wskazanie wielokryterialności związanej z analizowaniem tego zbioru danych. Jest to przykład rzeczywistego zbioru danych, dla oceny którego sama jakość klasyfikacji pozostaje niedoskonałą miarą. Co więcej, inne dostępne miary różnią się w zależności od zastosowanego algorytmu, a cele tych miar są zasadniczo różne. W ten sposób można stwierdzić, że algorytm CART będzie najlepszym rozwiązaniem w przypadku

analizy precyzji klasyfikacji dla rozmów telemarketingowych zakończonych sukcesem. W tym przypadku CART jest o ponad 1% lepszy od C4.5 i aż o niemal 15% lepszy od RT.

Jeśli natomiast czułość klasyfikacji jest czynnikiem wiodącym, to najlepiej wypada algorytm RT. Jest to ciekawa obserwacja, choć w dużej mierze taki rezultat związany jest z ogromną liczbą węzłów tak zbudowanego drzewa decyzyjnego. Algorytm TR jest lepszy od algorytmów CART i C4.5 odpowiednio o prawie 12% i ponad 6% pod względem miary czułości. Wyniki te można zaobserwować na rysunku 1, na którym zaprezentowano wartość analizowanych miar w zależności od zastosowanego algorytmu.



Rysunek 1. Wykres dla wartości analizowanych miar w zależności od zastosowanego algorytmu

Źródło: opracowanie własne.

W tym przypadku należy jednak rozważyć, czy stosowany klasyfikator może być tak duży. Co prawda klasyfikacja z zastosowaniem drzewa decyzyjnego jest stosunkowo szybką metodą, ale algorytm RT wymaga przeciętnie aż 1629 węzłów, kiedy algorytm CART (najlepszy w tym przypadku) składa się jedynie z 19 węzłów, a algorytm C4.5 ze 146 węzłów.

Podsumowanie

W artykule zaproponowano analizę zbioru danych zawierających informacje o kampanii telemarketingowej banku pod względem predykcji skuteczności rozmowy telefonicznej. Zastosowano w tym celu trzy algorytmy do konstruowania drzew decyzyjnych i przedstawiono uzyskane rezultaty.

Przeprowadzone eksperymenty potwierdzają, że drzewa decyzyjne są klasyfikatorami, które z powodzeniem można stosować do analizy tego rodzaju zbioru danych. Wyniki doświadczeń pozwalają dobrze określić predykcję przy zastosowaniu dokładności klasyfikacji. Nieco większy problem pojawia się przy analizie pod względem tylko jednej klasy decyzyjnej, która określa, czy rozmowa telemarketingowa przyniesie pozytywny skutek (klasa decyzyjna „Yes").

W tym przypadku zaproponowano ocenę algorytmu z zastosowaniem takich miar jakości, jak precyzja i czułość, a dodatkowo zaprezentowano macierze błędu. Obecnie wyniki konkretnych algorytmów różnią się w zależności od tego, jaki cel miałby podlegać predykcji.

W sytuacji, kiedy kampania telemarketingowa wymagałaby ograniczenia zasobów ludzkich poprzez zminimalizowanie liczby połączeń telefonicznych, ważniejsza staje się czułość (algorytm drzew losowych). Natomiast w przypadku, kiedy bank wspierałby metodę pozwalającą na wyznaczenie jak największej liczby pozytywnych rezultatów (z dopuszczeniem połączeń nieefektywnych), należałoby zastanowić się nad wybraniem algorytmu CART (najlepszy pod względem precyzji). Algorytm C4.5, który uzyskał najlepsze wyniki pod względem dokładności klasyfikacji, uśrednia wyniki z pozostałych miar.

Jak można zauważyć, analiza tego typu zbioru danych jest problemem wielokryterialnym. W przyszłości należałoby dokładnie zbadać ten zbiór danych w tym kontekście. Ponadto należy rozważyć zastosowanie algorytmów przybliżonych w celu wyznaczenia potencjalnych alternatywnych rozwiązań (klasyfikatorów).

Literatura

Boryczka, U., Kozak, J. (2014). *On-the-go adaptability in the new ant colony decision forest approach. In: Intelligent Information and Database Systems*. Intelligent Information and Database Systems – 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7–9, 2014, Proceedings, Part II. Springer International Publishing, 157–166.

Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D. (2013). *Weka manual for version 3-7-10*.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.

Elsalamony, H.A. (2014). Bank direct marketing analysis of data mining techniques. Network 5,0. *International Journal of Computer Applications (0975–8887), 85* (7), 12–22.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 29* (2), 119–127.

Koronacki, J., Ćwik, J. (2008). *Statystyczne systemy uczące się*. Warszawa: Exit.

Kozak, J. (2011). *Algorytmy mrowiskowe do konstruowania drzew decyzyjnych*. Nieopublikowana praca doktorska.

Kozak, J., Boryczka, U. (2013). Dynamic version of the acdt/acdf algorithm for h-bond data set analysis. *Computational Collective Intelligence. Technologies and Applications – 5th International Conference*. Craiova, Romania, September 11–13, Proceedings, 701–710.

Lim, T.-S., Loh, W.-Y., Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning, 40* (3), 203–228.

Loh, W.Y., Shih, Y.S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 815–840.

Moro, S., Cortez, P., Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22–31.

Moro, S., Laureano, R., Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. Conference-ESM'2011. *Eurosis*, 117–121.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1* (1), 81–106.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Quinlan, J.R. (1996). Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research, 4*, 77–90.

Rokach, L., Maimon, O. (2008). *Data Mining With Decision Trees: Theory And Applications.* River Edge, NJ, USA: World Scientific Publishing.

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications. Master's thesis*, Berlin: CASE Humboldt University.

Łapczyński, M. (2003). Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów. W: *Analiza satysfakcji i lojalności klientów. Zastosowania statystyki i data mining*, 93 –102. Kraków: AE w Krakowie.

ALGORITHMS FOR CONSTRUCTING DECISION TREES FOR PREDICTING THE EFFECTIVENESS OF THE BANK'S TELEMARKETING CAMPAIGN

KEYWORDS

decision trees, data analysis, telemarketing campaign

ABSTRACT

In this article we propose a novel approach for the generating transaction systems based on the technical analysis indicator - moving averages. Crossover of the moving average with the price chart is considered as a signal. Mechanism of setting the moving average period will be decreased in case of efficient trading. On the other hand, a couple of loss making trades leads to the increasing the moving average period. This will directly affect of decreasing number of trades. Such approach will be compared with the classical solutions based on crossover of two moving averages. Such mechanism will be presented as a system based on the procedural programming paradigm, in which stand-alone block codes are system functions. This will allow to easily expand some system functionalities without increasing code complexity.