STUDIA INFORMATICA 2016
Volume 37 Number 1 (123)

Łukasz PAŚKO, Galina SETLAK
Politechnika Rzeszowska, Zakład Informatyki

BADANIE JAKOŚCI PREDYKCYJNEJ SEGMENTACJI RYNKU

Streszczenie. Celem pracy jest ocena wyników predykcyjnej segmentacji rynku za pomocą narzędzi wykorzystywanych do badania jakości klasyfikatorów. Omawiana predykcyjna segmentacja rynku dotyczyła wyrobów gospodarstwa domowego. Przeprowadzono ją, wykorzystując klasyfikatory CART i CHAID. W pracy przedstawiono rezultaty oceny tych klasyfikatorów oraz wynikające z tego wnioski, dotyczące jakości segmentacji rynku.

Słowa kluczowe: klasyfikacja danych, ocena jakości klasyfikatorów, drzewa decyzyjne

AN EXAMINATION OF THE QUALITY OF PREDICTIVE MARKET SEGMENTATON

Summary. The aim of the paper is to assess the results of predictive market segmentation using methods of examination of classifiers' quality. The discussed predictive market segmentation was applied to household products. It was performed using CART and CHAID classifiers. The article contains the results of assessing the classifiers and the consequent conclusions on the quality of market segmentation.

Keywords: data classification, assessment of classifiers' quality, decision trees

1. Wstęp

Segmentacja rynku stanowi ważną część badań rynkowych. Pozwala znaleźć grupy podobnych do siebie nabywców lub produktów, zwane segmentami rynku. Prawidłowo przeprowadzona segmentacja może dać odpowiedź na wiele pytań istotnych z punktu widzenia przedsiębiorstwa działającego na danym rynku – może przyczynić się do lepszego poznania: potrzeb klientów, otoczenia przedsiębiorstwa, możliwości rynków zbytu, poziomu cen, efek-

tywności kanałów dystrybucji. Może być w końcu wskazówką do przyjęcia odpowiedniej strategii działania przedsiębiorstwa na danym rynku.

Obiekty należące do jednego segmentu charakteryzują się pewnym podobieństwem. W przypadku obiektów będących nabywcami lub potencjalnymi klientami przedsiębiorstwa, ich podobieństwo wyraża się zbliżonym zapotrzebowaniem, poziomem dochodów czy sposobem zakupu produktów. Gdy natomiast obiektem badań są produkty, to o ich wzajemnym podobieństwie decydują zbliżone cechy, porównywalna cena, poziom obsługi gwarancyjnej i posprzedażowej.

Literatura wyróżnia dwa podstawowe rodzaje segmentacji rynku:

- segmentacja opisowa – charakteryzuje się poszukiwaniem segmentów rynku w sposób nieukierunkowany, co spowodowane jest brakiem kryterium podziału rynku na segmenty;
- segmentacja predykcyjna – stosowana, gdy istnieje kryterium segmentacji [3, 7].

Do realizacji zadania segmentacji rynku można zastosować tradycyjne metody analizy danych, wykorzystujące metody statystyczne, takie jak analiza dyskryminacyjna czy metody regresji. Jednak coraz powszechniej stosowane są nowoczesne narzędzia informatyczne w połączeniu z technikami eksploracji danych, co ma się przyczynić do zwiększenia jakości wyników badań [1].

Z punktu widzenia analizy i eksploracji danych segmentacja może być realizowana z wykorzystaniem metod:

- grupowania danych – stosowane do segmentacji opisowej – cechy badanych obiektów są traktowane jako zmienne niezależne, natomiast zmienne zależne nie występują, przez co korzysta się tutaj np. z metod działających na zasadzie uczenia bez nauczyciela; wynikiem grupowania jest zidentyfikowanie skupisk podobnych do siebie obiektów stanowiących poszukiwane segmenty rynku [4, 8];
- klasyfikacji danych – w przypadku segmentacji predykcyjnej – każdy z badanych obiektów jest opisany nie tylko zestawem zmiennych niezależnych, ale ma także zmienną zależną, która przypisuje dany obiekt do jednej z klas; klasa odpowiada segmentowi rynku, a występowanie zmiennej zależnej umożliwia zastosowanie m.in. metod opartych na uczeniu z nauczycielem [6].

1.1. Dotychczasowe prace

Dotychczasowe prace skupiały się na analizach rynku zbytu wyrobów gospodarstwa domowego. Dysponowano zbiorem danych opracowanym na podstawie badań marketingowych. Zbiór danych zawiera cechy charakterystyczne 194 odkurzaczy będących obiektem tych badań. Każdy produkt opisany jest za pomocą dwunastu parametrów traktowanych w czasie analizy jako zmienne zależne (wejściowe). Dodatkowo zbiór zawiera zmienną nie-

zależną (wyjściową) CLASS, która na podstawie badań marketingowych przydziela każdy wyrób do jednego z czterech segmentów rynku: $\{m_1, m_2, m_3, m_4\}$.

Zrealizowane dotąd badania można podzielić na cztery etapy:

1. Opisowa segmentacja rynku. Jej celem było znalezienie nowych grup produktów i porównanie ich z predefiniowanymi segmentami opisanymi zmienną CLASS. Posłużono się siecią neuronową Kohonena, a wynikiem jej działania było odnalezienie czterech skupisk podobnych do siebie produktów, które potraktowano jako nowe segmenty rynku. Przynależność produktów do nowych segmentów zapisano w zbiorze danych w postaci drugiej zmiennej zależnej CLUSTER z wartościami $\{c_1, c_2, c_3, c_4\}$.
2. Predykcyjna segmentacja rynku względem zmiennej CLASS. Celem było utworzenie klasyfikatorów mogących przypisywać produkty do jednego z predefiniowanych segmentów $\{m_1, m_2, m_3, m_4\}$. Wykorzystano w tym miejscu drzewa decyzyjne CART (ang. *Classification and Regression Tree*) i CHAID (ang. *Chi-squared Automatic Interaction Detector*). Spośród kilkunastu utworzonych klasyfikatorów o różnym stopniu przycięcia do dalszej analizy wybrano dwa drzewa decyzyjne, które charakteryzowały się najmniejszą złożonością i najniższym kosztem sprawdzianu krzyżowego.
3. Predykcyjna segmentacja rynku względem zmiennej CLUSTER. Analogicznie do punktu 2, wybrano tutaj dwa najlepsze klasyfikatory CART i CHAID, które tym razem klasyfikują produkty, biorąc pod uwagę nowe segmenty rynku $\{c_1, c_2, c_3, c_4\}$.
4. Badanie jakości opisowej segmentacji rynku z punktu 1), zapisanej w zmiennej CLUSTER. Przeprowadzono go opierając się na wskaźnikach wykorzystywanych do badań jakości skupisk. Były to m.in. miary spójności klastrów, separacji międzyklastrowej, precyzja klasowa, entropia i jednorodność klastra. Otrzymane rezultaty zestawiono z wynikami otrzymanymi dla predefiniowanych segmentów rynku CLASS. Porównanie obu tych segmentacji pokazało, że sieć Kohonena odnalazła bardziej jednorodne grupy podobnych do siebie produktów.

Działania opisane w punktach 1-3 przedstawiono w pracy [12]. Natomiast punkt 4 jest tematem artykułu [9].

1.2. Zakres niniejszej pracy

W niniejszej pracy skupiono się na predykcyjnej segmentacji rynku. Celem omówionych tutaj badań jest ocena klasyfikatorów będących wynikiem punktów 2 i 3, opisanych w poprzedniej sekcji. Przeanalizowano wyniki klasyfikacji uzyskane dla obu zmiennych zależnych. Aby przedstawione analizy były bardziej przejrzyste, przyjęto stosować oznaczenia klasyfikatorów zaprezentowane w tabeli 1.

Każdy z czterech klasyfikatorów oceniano na podstawie kilkunastu znanych z literatury miarach jakości. Scharakteryzowano je krótko w sekcji 2 niniejszej pracy. Następnie porów-

nano uzyskane wyniki tych miar (sekcja 3) i podjęto próbę ogólnej oceny jakości zrealizowanej predykcyjnej segmentacji rynku (sekcja 4).

Tabela 1

Oznaczenia badanych klasyfikatorów

Oznaczenie	**Rodzaj**	**Zmienna zależna**	**Etykiety klas**
CART_M	drzewo decyzyjne CART	CLASS	$\{m_1, m_2, m_3, m_4\}$
CART_C	drzewo decyzyjne CART	CLUSTER	$\{c_1, c_2, c_3, c_4\}$
CHAID_M	drzewo decyzyjne CHAID	CLASS	$\{m_1, m_2, m_3, m_4\}$
CHAID_C	drzewo decyzyjne CHAID	CLUSTER	$\{c_1, c_2, c_3, c_4\}$,

2. Miary jakości segmentacji predykcyjnej

Ocenę jakości klasyfikatorów podzielono na dwie główne części:

- badanie ogólnej zdolności klasyfikacyjnej modelu dla wszystkich czterech klas łącznie, gdzie wykorzystano:
  - informację o rozmiarze drzewa decyzyjnego i wynik sprawdzianu krzyżowego,
  - macierz pomyłek,
- badanie zdolności klasyfikacyjnej każdej z poszczególnych klas z osobna, wykorzystując:
  - binarne macierze pomyłek,
  - wskaźniki liczbowe,
  - krzywe ROC.

Pierwszym z ocenianych parametrów jest rozmiar drzewa decyzyjnego. Wynika on bezpośrednio ze stopnia przycięcia drzewa, który odpowiada za wyeliminowanie przeuczenia klasyfikatora. Przyjęto, że bardziej preferowane będą mniej złożone klasyfikatory (o mniejszej liczbie węzłów decyzyjnych i liści). Złożoność modelu zestawiono z wynikami walidacji krzyżowej (ang. *k-fold cross-validation*). Jest ona użyteczna szczególnie w przypadku, gdy zbiór danych zawiera niewielką liczbę obiektów. Pozwala bowiem uniknąć wyodrębniania obiektów testowych ze zbioru danych, pomniejszając tym samym zbiór uczący. Zbiór danych dzieli się na *k* podzbiorów. *k*-1 podzbiorów wykorzystuje się jako dane uczące, natomiast pozostały podzbiór służy do testowania modelu. Procedurę powtarza się *k* razy, zmieniając każdorazowo podzbiór testowy i sprawdzając liczbę błędnych klasyfikacji. Uśredniony wynik wszystkich *k* powtórzeń powinien być jak najmniejszy.

Tabela 2

Binarna macierz pomyłek – ogólna postać

		Klasy przewidywane	
		pozytywna	negatywna
Klasy rzeczywiste	pozytywna	*TP*	*FN*
	negatywna	*FP*	*TN*

Z kolei macierz pomyłek pokazuje błędy klasyfikacyjne w rozbiciu na poszczególne klasy. Jest to kwadratowa macierz o wymiarach $n \times n$, gdzie n to liczba klas. Wiersze macierzy zawierają klasy poprawne (inaczej rzeczywiste lub oczekiwane), a kolumny – klasy przewidywane (decyzyjne). Wartość umieszczona w *i*-tym wierszu i *j*-tej kolumnie oznacza liczbę obiektów należących do *i*-tej klasy, a przypisanych przez klasyfikator do klasy *j*-tej.

W omawianych badaniach macierze pokazujące liczbę obiektów w każdej klasie mają wymiar 4 × 4. Jednak oprócz nich, postanowiono utworzyć tzw. macierze binarne, które są punktem wyjścia do drugiej części badania jakości. Pozwalają one skupić się na zdolności przewidywania tylko jednej klasy, nazywanej klasą pozytywną (inaczej wyróżnioną, relewantną). Pozostałe klasy traktowane są razem jako klasa negatywna (niewyróżniona, nierelewantna). Ogólną postać takiej macierzy przedstawia tabela 2.

Tabela 3

Wskaźniki wykorzystane do badania jakości klasyfikatorów

Symbol	**Popularne nazwy**	**Wzór**
Acc	*accuracy*, dokładność, poprawność frakcji	$Acc = \frac{TP+TN}{TP+TN+FP+FN}$
Err	*overall error rate*, łączny błąd klasyfikowania	$Err = \frac{FP+FN}{TP+TN+FP+FN}$
TPR	*true positives rate, recall, sensitivity, hit rate*, czułość, wrażliwość	$TPR = \frac{TP}{TP+FN}$
TNR	*true negatives rate, specificity*, swoistość	$TNR = \frac{TN}{TN+FP}$
PPV	*positive predictive value, precision*, precyzja	$PPV = \frac{TP}{TP+FP}$
NPV	*negative predictive value*, ujemna wartość predykcyjna	$NPV = \frac{TN}{TN+FN}$
FPR	*false positive rate, fall-out*, stopa fałszywych alarmów	$FPR = \frac{FP}{FP+TN} = 1-TNR$
FDR	*false discovery rate, fasle alarm ratio*, iloraz fałszywych alarmów	$FDR = \frac{FP}{FP+TP}$
FNR	*false negatives rate, miss rate*, współczynnik przeoczenia	$FNR = \frac{FN}{TP+FN} = 1-TPR$
MCC	*Matthew's correlation coefficient*	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(FN+TN)(FP+TN)}}$
*F*1	*F*1-*score*	$F1 = \frac{2 \times PPV \times TPR}{PPV + TPR}$
J	*Youden's J statistic*	$J = TPR + TNR - 1$

W polu *TP* (ang. *true positives*) umieszczana jest liczba poprawnie sklasyfikowanych obiektów należących do rzeczywistej klasy pozytywnej, natomiast pole *TN* (ang. *true negatives*) to poprawne klasyfikacje klasy negatywnej. Błędy klasyfikacyjne znajdują się w polach *FP* (ang. *false positives*) i *FN* (ang. *false negatives*). Oznaczają one odpowiednio: sklasyfikowanie obiektów z klasy pozytywnej do negatywnej (tzw. błędy pierwszego rodzaju) oraz przypisanie przypadków z klasy negatywnej do pozytywnej (tzw. błędy drugiego rodzaju).

Na podstawie binarnych macierzy pomyłek obliczono kilkanaście znanych z literatury wskaźników liczbowych. Wszystkie wykorzystane wskaźniki zaprezentowano w tabeli 3. Są one przedstawione i omówione m.in. w pracach [2, 5, 10, 11, 13].

Ostatnią metodą oceny drzew decyzyjnych była analiza ROC. Krzywa ROC prezentuje zależność pomiędzy wrażliwością klasyfikatora, czyli odsetkiem wyników prawdziwie dodatnich (*TPR* – oś pionowa wykresu ROC), a 1-specyficznością, czyli odsetkiem wyników fałszywie dodatnich (*FPR* – oś pozioma). Analiza ROC wykorzystywana jest najczęściej do przedstawienia, w jaki sposób zmiana wartości progowej klasyfikatora (ang. *threshold*) wpływa na jego zdolność klasyfikowania. Dzięki analizie ROC możliwy jest dobór optymalnej wartości progowej, zwanej także punktem odcięcia (ang. *cut-off*). Patrząc na krzywą ROC tylko w tym kontekście, przeprowadzanie analizy ROC miałoby sens jedynie dla modelu, który na wyjściu daje pewną wartość liczbową określającą stopień przynależności do klasy (ang. *scoring*), przykładowo: wartość funkcji aktywacji neuronu wyjściowego perceptronu czy też wartość prawdopodobieństwa wyznaczana przez naiwny klasyfikator bayesowski [5, 10].

Jednak krzywą ROC można wykorzystać również jako miernik jakości klasyfikatora, wyznaczając pole powierzchni pod krzywą (*AUC* – ang. *area under curve*). W taki właśnie sposób wykorzystano analizę ROC w niniejszej pracy.

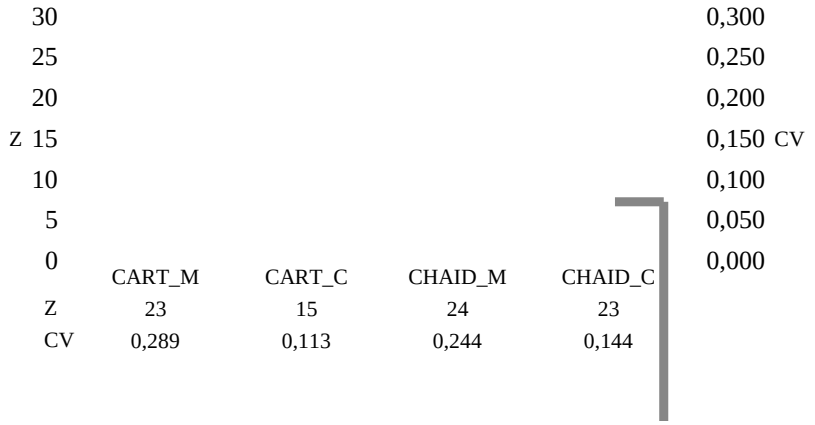
3. Wyniki analiz

W niniejszej sekcji skupiono się najpierw na ogólnym badaniu zdolności klasyfikacyjnej modelu dla wszystkich czterech klas łącznie. W drugiej części opisano ocenę klasyfikowania każdej z czterech klas z osobna. Wyniki pierwszej i drugiej części omówiono dla wszystkich czterech modeli.

3.1. Ogólna ocena klasyfikatorów

Pierwszy krok oceny jakości klasyfikatorów polegał na porównaniu rozmiaru drzewa decyzyjnego z wynikiem *k*-krotnego sprawdzianu krzyżowego. Sprawdzian krzyżowy wykonano dla $k = 10$. Rezultaty zaprezentowano na rysunku 1. Rozmiar drzewa decyzyjnego wyra-

żono przez jego złożoność Z obliczaną jako suma liczby węzłów decyzyjnych i liści klasyfikatora. Walidację krzyżową zrealizowano z wykorzystaniem oprogramowania STATISTICA Data Miner. W zastosowanym oprogramowaniu wynik walidacji krzyżowej dla modeli CART jest nazywany kosztem sprawdzianu krzyżowego, natomiast dla drzew CHAID nosi on nazwę ryzyka estymacji.



Rys. 1. Złożoność klasyfikatorów i wyniki sprawdzianu krzyżowego
Fig. 1. Complexity of the classifiers and results of cross-validation

Porównując wartości wyznaczone dla modeli CART, można stwierdzić, że lepszą zdolność do predykcji będzie mieć drzewo CART_C. Sprawdzian krzyżowy daje w jego przypadku rezultat ponaddwukrotnie mniejszy niż dla modelu CART_M. Poza tym występuje znaczna różnica w złożoności obu modeli przemawiająca na korzyść prostszego drzewa decyzyjnego. Wynika z tego, że algorytm CART łatwiej znalazł zależności występujące pomiędzy cechami klasyfikowanych obiektów, analizując dane ze względu na zmienną zależną CLUSTER.

W przypadku drzew CHAID złożoność obydwu modeli jest bardzo podobna. Natomiast ryzyko estymacji również przemawia na korzyść drzewa utworzonego dla zmiennej CLUSTER. Potwierdza to sytuację zaobserwowaną w modelach CART – reguły opisujące powiązania między cechami obiektów a ich klasami są łatwiej wychwytywane przez klasyfikator działający dla segmentów rynku znalezionych przez sieć Kohonena niż dla segmentów będących wynikiem badań rynkowych.

Ostatnim elementem ogólnej analizy klasyfikatorów są ich macierze pomyłek. Utworzono je dla wszystkich czterech klasyfikatorów. Jednak w niniejszej pracy zaprezentowano tylko dwie macierze, które najbardziej różniły się pod względem błędów klasyfikacyjnych (tabela 4). Najmniejszą liczbę niepoprawnych klasyfikacji (16) odnotowano dla modelu CART_C (lewa część tabeli 4). Z kolei najwięcej błędnych klasyfikacji popełnia drzewo CHAID_M (34). W macierzach pozostałych dwóch klasyfikatorów wystąpiło 31 pomyłek w przypadku CART_M oraz 17 pomyłek dla CHAID_C. Biorąc pod uwagę te wartości, można zauważyć, że to zmienna zależna odpowiada za większą liczbę błędów, a nie dobór algorytmu tworzące-

go drzewo decyzyjne – zarówno w modelu CART, jak i CHAD wynik predykcji był gorszy, klasyfikując ze względu na zmienną CLASS. Te spostrzeżenia odpowiadają wynikom omówionej powyżej walidacji krzyżowej.

Tabela 4

Macierz pomyłek dla modeli CART_C i CHAID_M

	Klasa przewidywana					Klasa przewidywana			
Klasa rzeczywista	c_1	c_2	c_3	c_4	Klasa rzeczywista	m_1	m_2	m_3	m_4
c_1	34	0	2	0	m_1	28	5	1	0
c_2	0	37	6	0	m_2	6	27	9	0
c_3	0	3	77	1	m_3	4	7	60	1
c_4	0	3	1	30	m_4	0	0	1	45

3.2. Szczegółowa ocena klasyfikatorów

Druga część badania skupia się na ocenie zdolności do klasyfikowania poszczególnych segmentów rynku z osobna. Badanie wykonano dla wszystkich czterech drzew decyzyjnych. Analiza polegała na sporządzeniu czterech binarnych macierzy pomyłek dla każdego klasyfikatora. W macierzy binarnej analizowana w danym momencie klasa traktowana była jako pozytywna (wyróżniona), natomiast pozostałe klasy tworzyły razem klasę negatywną (niewyróżnioną). Na podstawie tak sporządzonych macierzy binarnych obliczono dla każdej z nich wartości dwunastu wskaźników świadczących o jakości klasyfikatora. W tabeli 5 zebrano wyniki dla dwóch klasyfikatorów, które najbardziej się różniły pod względem ogólnej oceny jakości z sekcji 3.1.

Aby łatwiej ocenić powyższe rezultaty, tabelę 5 podzielono na dwie części. Górna część (od *Acc* do *J*) zawiera wskaźniki, których wartość oczekiwana, rozumiana jako wynik uzyskany dla klasyfikatora niepopełniającego błędów klasyfikacyjnych, wynosi 1. Pozostałe cztery wskaźniki powinny osiągać jak najmniejsze wartości – w przypadku bezbłędnego klasyfikatora wynik będzie równy 0. Wartości wszystkich wskaźników, z wyjątkiem *MCC* i *J*, mieszczą się w zakresie [0; 1]. Wynik zbliżony do 0,5 oznacza, że klasyfikator działa jak klasyfikator losowy (zastosowanie badanego modelu daje taki sam skutek jak losowe przydzielanie obiektów do klasy c^+ bądź c^-). Gdy wynik oddala się jeszcze bardziej od wartości oczekiwanej, wówczas badany klasyfikator daje więcej błędów predykcyjnych niż poprawnych klasyfikacji. W przypadku *MCC* oraz *J* zakres to [–1; 1], a klasyfikator losowy przyjmuje wartość 0.

Pierwszy wskaźnik (dokładność – *Acc*) pokazuje, że model CART_C z większym prawdopodobieństwem przydziela obiekty do tej klasy, do której rzeczywiście one należą. Wyjątkiem jest tylko czwarty segment, dla którego CHAID_M ma minimalnie lepszą wartość. Różnice pomiędzy modelami widać szczególnie dobrze na przykładzie drugiego i trzeciego segmentu rynku. Podobną tendencję ujawnia czułość klasyfikatorów (*TPR*). Zdolność wy-

krywania obiektów z klasy wyróżnionej jest większa w przypadku CART_M, z wyjątkiem czwartego segmentu – dla tego segmentu rynku CART_C popełnia więcej błędów drugiego rodzaju niż model CHAID_M. Patrząc na swoistość (*TNR*), widać z kolei, że wyniki dla każdej klasy są zbliżone. Oba modele mają więc podobną zdolność wykrywania obiektów nienależących do klasy wyróżnionej i popełniają mało błędów pierwszego rodzaju.

Tabela 5

Wartości wskaźników dla modeli CART_C i CHAID_M

	CART_C				CHAID_M			
	klasa wyróżniona							
wskaźniki	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
Acc	,99	,94	,93	,97	,92	,86	,88	,99
TPR	,94	,86	,95	,88	,82	,64	,83	,98
TNR	1,	,96	,92	,99	,94	,92	,91	,99
PPV	1,	,86	,90	,97	,74	,69	,85	,98
NPV	,99	,96	,96	,98	,96	,90	,90	,99
MCC	,97	,82	,86	,91	,73	,58	,75	,97
F1	,97	,86	,92	,92	,78	,67	,84	,98
J	,94	,82	,87	,88	,76	,56	,74	,97
Err	,01	,06	,07	,03	,08	,14	,12	,01
FPR	,00	,04	,08	,01	,06	,08	,09	,01
FDR	,00	,14	,10	,03	,26	,31	,15	,02
FNR	,06	,14	,05	,12	,18	,36	,17	,02

Precyzja (*PPV*) drzewa CART_C jest znacznie większa od CHAID_M dla pierwszej i drugiej klasy. Tak więc prawdopodobieństwo tego, że obiekt sklasyfikowany za pomocą CART_C jako pierwszy lub drugi segment faktycznie należy do pierwszego lub drugiego segmentu, jest znacznie większe od prawdopodobieństwa takiej klasyfikacji przy użyciu CHAID_M. Natomiast ujemna wartość predykcyjna (*NPV*), czyli prawdopodobieństwo przynależności obiektu uznawanego przez klasyfikator za niewyróżniony do rzeczywistej klasy niewyróżnionej, jest zbliżone, patrząc na wszystkie klasy obu modeli.

Miara *MCC* jest współczynnikiem korelacji pomiędzy klasami rzeczywistymi a przewidywanymi. Bierze on pod uwagę całą macierz pomyłek, a nie tylko pewien jej wycinek, jak robią to opisane wcześniej wskaźniki. Dlatego *MCC* jest dobrą miarą do ogólnego spojrzenia na jakość klasyfikacji. Wyniki obu porównywanych modeli wskazują, że CART_C jest lepszym klasyfikatorem dla pierwszych trzech segmentów rynku, zaś dla czwartego segmentu model CHAID_M osiąga lepszy wynik. Dwie ostatnie miary górnej części tabeli są wyznaczane na podstawie wcześniej omówionych wskaźników. Miara *F*1 jest średnią harmoniczną precyzji i czułości modelu, natomiast statystyka *J* Joudena to suma czułości i swoistości pomniejszona o 1. Obydwa wskaźniki mierzą zatem ogólną skuteczność dyskryminacyjną klasyfikatora. Rezultaty *F*1 i *J* dla obu modeli są analogiczne do wyników *MCC*.

Pozostałe wskaźniki powinny przyjmować jak najniższe wartości. Ogólny błąd klasyfikatora (*Err*) dla trzech pierwszych segmentów jest korzystniejszy w przypadku stosowania mo-

delu CART_C. Podobną zależność wskazują także *FPR* (prawdopodobieństwo fałszywych alarmów, czyli obiektów niepoprawnie przypisanych do klasy wyróżnionej, wśród wszystkich obiektów rzeczywiście niewyróżnionych), *FDR* (prawdopodobieństwo fałszywych alarmów wśród wszystkich obiektów uznanych przez klasyfikator za wyróżnione) oraz *FNR* (prawdopodobieństwo przeoczenia obiektów wyróżnionych, czyli przypisania ich przez klasyfikator do klasy niewyróżnionej). Jedynie w przypadku czwartego segmentu rynku, skuteczność CHAID_M jest porównywalna w stosunku do CART_C lub minimalnie od niego lepsza.

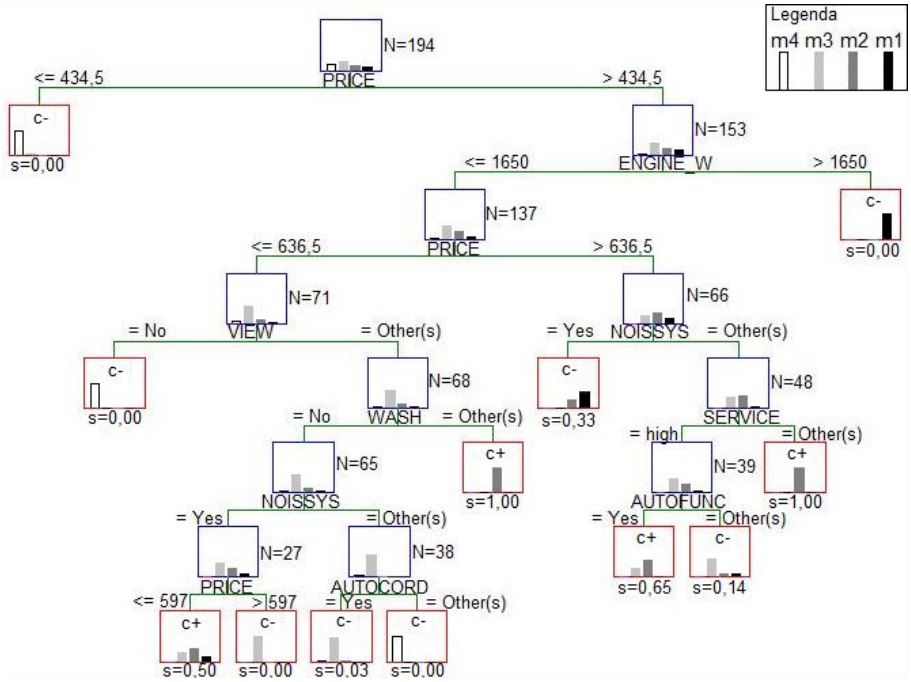
Wszystkie miary dla pozostałych dwóch klasyfikatorów osiągały podobne rezultaty. Dla modelu CHAID_C wartości były bardzo zbliżone do CART_C, natomiast w przypadku modelu CART_M wyniki były analogiczne do uzyskanych dla CHAID_M. Potwierdza to sytuację zaobserwowaną w sekcji 3.1 – predykcje dla segmentów rynku ze zmiennej CLASS obarczone są większym błędem niż predykcje segmentów znalezionych przez sieć Kohonena.

Ostatnia metoda oceny polegała na sporządzeniu krzywych ROC na podstawie macierzy binarnych i obliczeniu *AUC*. Modele omawianych drzew decyzyjnych nie dają na wyjściu wartości liczbowych określających stopień przynależności obiektów do klas (ang. *scoring*). Uniemożliwia to wyznaczenie krzywej ROC bezpośrednio na podstawie wartości wyjściowych z klasyfikatora. Dlatego w celu sporządzenia krzywej ROC posłużono się metodą opisaną w poniższych krokach:

1) oznacz jedną klasę c_i spośród n wszystkich klas jako wyróżnioną (c^+),
2) oznacz pozostałe klasy c_j jako niewyróżnione (c^-), przy czym $j = 1,...,n-1$ oraz $j \neq i$,
3) dla każdego liścia L klasyfikatora:
   a) oblicz iloraz $s_L = \dfrac{N_L(c^+)}{N_L(c^+) + N_L(c^-)}$, gdzie $N_L(c^+)$ to liczba obiektów ze zbioru uczącego klasyfikowanych przez liść L do klasy wyróżnionej, natomiast $N_L(c^-)$ to liczba obiektów przydzielanych przez liść L do którejkolwiek z klas niewyróżnionych,
   b) jeśli $N_L(c_i) > \max_j \{N_L(c_j)\}$, to przypisz do L etykietę c^+,
   c) w przeciwnym razie przypisz do L etykietę c^-,
4) dla każdego obiektu ze zbioru danych:
   a) uruchom klasyfikator,
   b) przypisz obiekt do klasy c^+ lub c^-,
   c) przypisz obiektowi wartość s_L tego liścia L, który dokonał klasyfikacji obiektu,
5) sporządź wykres krzywej ROC za pomocą dowolnej metody wykorzystywanej do tego celu, traktując wartości s_L otrzymane na wyjściu klasyfikatora jako scoring,
6) oblicz *AUC* dla uzyskanej krzywej.

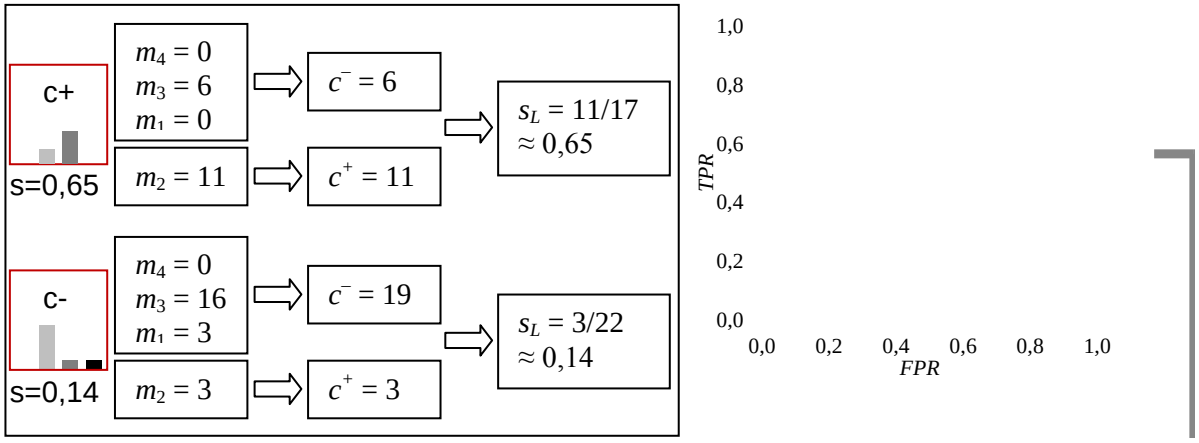
Rezultaty metody opisanej w powyższym algorytmie przedstawiono na przykładzie klasyfikatora CART_M w sytuacji, gdy klasa wyróżniona to m_2, zaś klasy niewyróżnione to m_1, m_3, m_4.

Drzewo decyzyjne CART_M dla tej sytuacji zaprezentowano na rysunku 2. W węzłach decyzyjnych i liściach umieszczono histogramy pokazujące liczbę obiektów należących do klas oryginalnych. Obok każdego liścia umieszczono również wartość ilorazu s_L, a obok węzłów decyzyjnych – liczność rozpatrywanego w danym węźle podzbioru obiektów uczących. W przypadku gdy w danym liściu klasą większościową jest m_2, wówczas taki liść zaliczany jest do klasy c^+. Jeśli klasą większościową jest którakolwiek z klas niewyróżnionych, to liść otrzymuje etykietę c^-.



Rys. 2. Drzewo CART_M dla wyróżnionej klasy m_2
Fig. 2. CART_M tree for class m_2 as the positive

Wynikiem działania klasyfikatora jest przypisanie wszystkich obiektów ze zbioru danych do klasy c^+ lub c^- wraz z informacją o scoringu każdego obiektu. Mając tak przygotowane dane, wyznacza się krzywą ROC. Uzyskana w tym przypadku krzywa przedstawiona jest na rysunku 3. Ostatni krok to obliczenie pola powierzchni pod krzywą, które dla opisanego modelu CART_M i wyróżnionej klasy m_2 wynosi 0,93. Obok wykresu ROC przedstawiono graficznie sposób wyznaczania scoringu dla dwóch wybranych liści z rysunku 2.



Rys. 3. Sposób obliczania scoringu i krzywa ROC dla drzewa z rys. 2
Fig. 3. Calculation of scoring and ROC curve for the tree from fig. 2

W analogiczny sposób sporządzono krzywe ROC i obliczono miary *AUC* dla każdej klasy wyróżnionej we wszystkich czterech badanych klasyfikatorach. Wyniki zebrano w tabeli 6. Również te rezultaty wskazują bardzo podobną zdolność klasyfikowania czwartego segmentu rynku (c_4 i m_4) za pomocą każdego z modeli. Większe różnice występują dla pozostałych segmentów, wskazując przy tym na lepszą jakość modeli CART_C i CHAID_C.

Tabela 6

Wartości *AUC* badanych klasyfikatorów

	klasa wyróżniona							
	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
CART	,99	,95	,97	,98	,97	,93	,95	,99
CHAID	,99	,99	,98	,99	,95	,87	,93	,99

4. Podsumowanie

W niniejszej pracy badano jakość predykcyjnej segmentacji rynku realizowanej z wykorzystaniem drzew decyzyjnych CART i CHAID. Analizowano dwa rodzaje segmentów rynku: pochodzące z badań rynkowych (CLASS) oraz ustalone za pomocą grupowania danych siecią Kohonena (CLUSTER). Pierwsza część badania jakości polegała na ogólnym spojrzeniu na wyniki generowane za pomocą klasyfikatorów. Uzyskane rezultaty pokazują, że jakość klasyfikatorów zależy od zmiennej zależnej, dla której wykonywano klasyfikację. Wtedy gdy zmienną zależną była CLASS, wówczas predykcje modeli były obarczone większym błędem niż w przypadku, gdy klasyfikowano względem zmiennej CLUSTER. Przemawia to na korzyść segmentacji uzyskanej za pomocą sieci Kohonena.

W drugiej części analizowano zdolność predykcji każdego z segmentów rynku z osobna. Rezultaty pokazywały tutaj, że klasyfikatory popełniają najmniej błędów dla skrajnych segmentów rynku: pierwszego (produkty najbardziej zaawansowane) oraz czwartego (produkty

najgorsze). Produkty z klas środkowych okazały się trudniejsze do rozpoznania. Poza tym, w tym miejscu również można było zauważyć, że klasyfikatory działające na zmiennej CLUSTER dają lepsze rezultaty dla każdego segmentu rynku. Wyjątkiem jest tylko segment czwarty – zarówno dla zmiennej CLASS, jak i CLUSTER badane modele mają podobną zdolność klasyfikowania tego segmentu. Tak więc należy uznać, że jakość segmentacji predykcyjnej utworzonej na podstawie grup sieci Kohonena jest lepsza od jakości segmentacji pochodzącej z badań rynkowych. Wyniki tych segmentacji są zbliżone jedynie dla segmentu najmniej zaawansowanych produktów.

BIBLIOGRAFIA

1. Cios K., Pedrycz W., Świniarski R.: Data mining methods for knowledge Discovery. Kluwer, Norwell MA 1998.
2. Costa E.P., Lorena A.C., Carvalho A.C.P.L.F., Freitas A.A.: A review of performance evaluation measures for hierarchical classifiers. Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop, AAAI Press, 2007, s. 182÷196.
3. Dolnicar S.: Using cluster analysis for market segmentation – typical misconceptions, established methodological weaknesses and some recommendations for improvement. Australasian Journal of Market Research, Vol. 11(2), 2003, s. 5÷12.
4. Everitt B.S., Landau S., Leese M.: Cluster analysis. Wiley Publishing, Nowy Jork 2009.
5. Fawcelt T.: An introduction to ROC analysis. Pattern Recognition Letters, Vol. 27, Issue 8, Elsevier, New York 2006, s. 861÷874.
6. Gordon A.D.: Classification, 2nd edition. Chapman & Hall/CRC Press, 1999.
7. Migut G.: Zastosowanie technik analizy skupień i drzew decyzyjnych do segmentacji rynku. Materiały Seminarium StatSoft „Zastosowanie nowoczesnej analizy danych w marketingu i badaniach rynku", Kraków 2010.
8. Nowak-Brzezińska A., Xięski T.: Grupowanie danych złożonych. Zeszyty Naukowe Politechniki Śląskiej, Seria Informatyka, Vol. 32, No. 2A(96), s. 391÷401.
9. Paśko Ł., Setlak G.: Ocena segmentacji rynku za pomocą miar jakości grupowania danych. Zeszyty Naukowe Politechniki Śląskiej, Seria Informatyka, Vol. 35, No. 2(116), Gliwice 2014, s. 157÷173.
10. Powers D.M.W.: Evaluation: from precision, recall and F-score to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, Vol. 2, 2011, s. 37÷63.

11. Provost F., Fawcett T., Kohavi R.: The case against accuracy estimation for comparing classifiers. Proceedings of the ICML-98. Morgan Kaufmann, San Francisco 1998, s. 445÷453.
12. Setlak G., Paśko Ł.: Zastosowanie metod eksploracji danych do segmentacji rynków. Zeszyty Naukowe Politechniki Śląskiej, Seria Informatyka, Vol. 34, No. 2A(111), Gliwice 2013, s. 311÷323.
13. Sokolova M., Lapalme G.: A systematic analysis of performance measures for classification tasks. Information Processing and Management, Vol. 45, Issue 4, Elsevier, 2009, s. 427÷437.

Abstract

This paper investigated the quality of predictive market segmentation implemented using CART and CHAID decision trees. Two types of market segments were analyzed: segments that came from marketing research, and determined using Kohonen neural network. The first part of the study is to overall look at the classification results (section 3.1). This section takes into account the complexity of the decision trees, the results of cross-validation, and the confusion matrices. In the second part (section 3.2) the ability of prediction of each segment separately is analyzed. For this purpose binary confusion matrices were prepared, and several quality indicators were calculated. These measures showed that classifiers make the fewest mistakes for peripheral market segments: the first, which collects the most advanced products, and the last with the worst products. Moreover, the classifiers operating on Kohonen market segments produce better predictive results than the models, which classify the segments from marketing research. It can be concluded that the quality of predictive segmentation created for Kohonen market segments is better than the second segmentation.

Adresy

Łukasz PAŚKO: Politechnika Rzeszowska, Zakład Informatyki, al. Powstańców Warszawy 8, 35-959 Rzeszów, Polska, lpasko@prz.edu.pl.
Galina SETLAK: Politechnika Rzeszowska, Zakład Informatyki, al. Powstańców Warszawy 8, 35-959 Rzeszów, Polska, gsetlak@prz.edu.pl.