STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Galina SETLAK, Łukasz PAŚKO
Politechnika Rzeszowska, Zakład Informatyki

ZASTOSOWANIE METOD EKSPLORACJI DANYCH DO SEGMENTACJI RYNKÓW

Streszczenie. Celem niniejszego artykułu są przedstawienie i ocena możliwości wykorzystania metod eksploracji danych do segmentacji rynków zbytu. Przedstawiono segmentacje opisową i predykcyjną oraz przeanalizowano wyniki rozwiązywania zadań klasyfikacji i grupowania danych za pomocą sieci neuronowych Kohonena oraz drzew klasyfikacyjnych CART i CHAID. W pracy wykorzystano dane dotyczące rynków zbytu przedsiębiorstwa produkującego wyroby gospodarstwa domowego.

Słowa kluczowe: eksploracja danych, grupowanie danych, klasyfikacja danych, sieci neuronowe Kohonena, drzewa klasyfikacyjne

APPLICATION OF DATA MINING METHODS ON MARKETS SEGMENTATION

Abstract. The purpose of this paper is to present and evaluate the possibility of using data mining methods in the market segmentation process. In the paper the descriptive and predictive segmentation were presented and the results of classification and clustering data were analyzed. To carry out the analysis were used following methods: Kohonen neural networks, CART and CHAID. The analysis concerns the manufacturing company producing household products.

Key words: data mining, data clustering, data classification, Kohonen neural networks, classification trees

1. Wprowadzenie

Badania rynkowe i marketingowe przygotowują dla menedżera bardzo ważną wiedzę o potencjalnych możliwościach rynków zbytu, charakterystykach zapotrzebowań na nich, o tendencjach ustalania cen i efektywności reklamy [9]. Ponadto, dają możliwość poznania charakterystyk rynku, kanałów dystrybucji i strategii stymulacji zbytu. Segmentacja rynku

może być wykonana zarówno w badaniach rynkowych, jak i marketingowych i polega na podziale rynku w celu wyodrębnienia grup nabywców o podobnych cechach, dochodach, potrzebach, reakcji na reklamę, preferencjach odnośnie do sposobu zakupu itp. Po dokonaniu segmentacji firma musi wybrać segment, w którym chce sprzedawać swoje towary.

Tradycyjnie do analizy danych rynkowych i marketingowych wykorzystywane były metody statystyczne, takie jak: metody regresji i korelacji, analiza dyskryminacyjna oraz analiza czynnikowa. W celu zwiększenia efektywności wyników badań, do ich analizy coraz powszechniej są wykorzystywane zaawansowane, współczesne narzędzia informatyczne oraz metody sztucznej inteligencji [11, 12] i eksploracji danych (ang. *data mining*) [2, 3, 8]. Eksploracja danych polega na efektywnym znajdowaniu ukrytych dla człowieka prawidłowości lub nieznanych dotychczas zależności i związków pomiędzy danymi w dużych zbiorach danych, które mogą być wykorzystane jako wiedza dla menedżera przy podejmowaniu decyzji.

Celem pracy są przedstawienie i ocena możliwości wykorzystania metod eksploracji danych do wspomagania decyzji w procesie segmentacji rynków zbytu. Przedstawiono w niej wyniki segmentacji opisowej rynku za pomocą sieci neuronowych Kohonena oraz segmentacji predykcyjnej z zastosowaniem drzew klasyfikacyjnych, w tym narzędzi CART (ang. *Classification and Regression Tree*) i automatycznego detektora interakcji za pomocą chi-kwadrat CHAID (ang. *Chi-squared Automatic Interaction Detector*). Badania realizowano w celu wspomagania segmentacji rynku zbytu wyrobów dla przedsiębiorstwa produkującego sprzęt gospodarstwa domowego.

2. Podstawy teoretyczne

Segmentacja rynku polega na jego podzieleniu na mniejsze części, zwane segmentami, które różnią się między sobą oczekiwaniami klientów względem produktu, sposobem zakupu lub innymi kryteriami. Głównym celem segmentacji jest analiza potrzeb klientów, które tworzą rynek. Drugim celem segmentacji jest pozycjonowanie produktu, czyli nadanie mu, w odbiorze klientów, pewnych specyficznych atutów, wyróżniających produkt względem konkurentów i innych segmentów [8, 9]. W procesie segmentacji rynku muszą być przede wszystkim określone kryteria segmentacji, w zależności od których dobierana jest odpowiednia metoda grupowania. Kolejnym ważnym krokiem jest ustalenie optymalnej liczby segmentów, o ile jest to możliwe.

W literaturze znane są dwa rodzaje segmentacji: segmentacja opisowa i segmentacja predykcyjna. Segmentacja opisowa rynku jest stosowana w sytuacjach, gdy brak jest kryterium, które ukierunkowałoby poszukiwanie grup, wszystkie zmienne (cechy opisujące obiekty) są zmiennymi niezależnymi oraz brak jest informacji pozwalającej na stosowanie metody ucze-

nia z nauczycielem i możemy wykorzystać tylko jedną z metod tak zwanej nieukierunkowanej eksploracji danych (uczenie bez nauczyciela) [8]. Wynikiem tak przeprowadzonej segmentacji jest podział konsumentów na grupy oraz dobranie odpowiednich kryteriów tego podziału. Do realizacji opisowej segmentacji rynku można zastosować dobrze znane metody, takie jak metoda aglomeracyjna oraz sztuczne sieci neuronowe Kohonena.

Segmentacja predykcyjna stosowana jest w sytuacjach, gdy istnieje możliwość określenia kryterium segmentacji, występują zmienne zależne oraz można wykorzystać metody ukierunkowanej eksploracji danych, oparte na uczeniu z nauczycielem. Segmentację predykcyjną można więc wykonać za pomocą metod i narzędzi do rozwiązywania zadań klasyfikacji i grupowania, należących do podstawowych technik eksploracji danych, w tym również sieci neuronowych (nauczanie z nauczycielem) oraz drzew decyzyjnych, takich jak: CART, CHAID, a także drzew wzmacnianych (ang. *boosted trees*).

Klasyfikacja i grupowanie są podstawowymi, ważniejszymi metodami powszechnie stosowanymi w analizie danych. Przy czym klasyfikacja oznacza przyporządkowanie obiektu na podstawie wybranych cech charakterystycznych do jednej z klas wzorcowych. O klasyfikacji mówimy wtedy, jeżeli klasy, do których będzie rozdzielany zbiór wejściowy, zostaną zdefiniowane przed procesem podziału [1, 14, 15].

Grupowanie (ang. clustering) jest działaniem pierwotnym w stosunku do klasyfikacji, ponieważ prowadzi do zdefiniowania klas. O grupowaniu mówimy wtedy, gdy przed podziałem zbioru wejściowego nie jest określona liczba grup oraz nie są znane przedziały wartości cech charakterystycznych. Przez grupowanie rozumiemy kojarzenie obiektów podobnych i rozdzielanie różnych [1, 2, 14]. Grupowanie oznacza podział zbioru elementów na podzbiory, na podstawie cech charakterystycznych wykrytych podczas procesu podziału.

W niniejszych badaniach do analizy danych rynkowych i wspomagania segmentacji rynku zostały wykorzystane podstawowe narzędzia eksploracji danych, takie jak sztuczne sieci neuronowe Kohonena i drzewa decyzyjne CART i CHAID.

3. Eksperyment

W badaniach dotyczących segmentacji rynku zbytu wyrobów gospodarstwa domowego wykorzystano zbiory danych opracowane na podstawie danych ankietowych, uzyskanych w wyniku badań marketingowych i rynkowych dla przedsiębiorstwa produkcyjnego, w latach 2003-2005. Zbiory danych zawierają opis cech charakterystycznych odkurzaczy, które są w niniejszych badaniach obiektem badań. Właściwie realizowany jest pierwszy etap segmentacji, polegający na podziale rynku zbytu na kilka segmentów, czyli rozwiązywane są zadania

grupowania i klasyfikacji. W tabeli 1 są przedstawione parametry wejściowe, będące cechami charakterystycznymi wyrobów.

Tabela 1

Cechy charakterystyczne wyrobów

Atrybut	Opis	Typ danych
ENGINE_W	Moc silnika, W	numeryczny
PRICE	Cena	numeryczny
FILTR_SYS	Zaawansowany system filtrujący	{tak, nie}
AUTOFUNC	Automatyczna kontrola mocy	{tak, nie}
AUTOCORD	Zwijacz przewodu zasilającego	{tak, nie}
SPD_CTRL	Elektroniczna regulacja siły ssania	{tak, nie}
NOISSYS	System tłumienia hałasu	{tak, nie}
WASH	Opcja czyszczenia na mokro	{tak, nie}
VIEW	Styl i design	{tak, nie}
FEATURE	Dodatkowe cechy	{tak, nie}
BRAND	Marka	{tak, nie}
SERVICE	Obsługa posprzedażowa	{niski, średni, wysoki}

Jako parametr wyjściowy w przygotowanym zbiorze danych występuje jeden z segmentów rynku, który zostanie wybrany w wyniku wykonania zadania klasyfikacji, oznaczony w zbiorze uczącym jako CLASS.

3.1. Segmentacja opisowa z wykorzystaniem sieci neuronowych Kohonena

Segmentację rynku w celu grupowania danych najpierw zrealizowano za pomocą sieci neuronowych Kohonena, z wykorzystaniem pakietu programowego Statistica Neural Networks.

Sieć neuronowa Kohonena nazwana jest przez jej twórcę – Teuvo Kohonena [5] – samoorganizującym się odwzorowaniem *(Self-Organizing Maps – SOM)* lub samoorganizującym się odwzorowaniem cech *(Self-Organizing Feature Maps – SOFM)*. Jest to najbardziej popularna sieć neuronowa, reprezentująca rodzaj sieci samoorganizujących się, które są tak nazywane ze względu na sposób nauczania – nauczanie bez nauczyciela (*unsupervised learning*). Sieć Kohonena zawiera M neuronów, tworzących na płaszczyźnie prostokątną siatkę. Sygnały wejściowe są podawane na wejścia wszystkich neuronów w sieci. Warstwa wyjściowa pełni rolę obliczeniową i jednocześnie prezentuje wyniki. Neurony tworzą tutaj mapę topologiczną, dzięki której można obserwować znalezione w zbiorze danych skupiska.

Podstawową metodą uczenia się sieci samoorganizujących się, w tym sieci Kohonena, jest metoda uczenia konkurencyjnego [5, 6]. Celem tego rodzaju sieci jest grupowanie lub klasyfikowanie wzorców wejściowych. Odbywa się to zgodnie z zasadą, że podobne sygnały wejściowe wzbudzają te same jednostki wyjściowe sieci neuronowej. Grupy są definiowane na podstawie korelacji danych wejściowych.

Przyjęto, że atrybut CLASS nie będzie brany pod uwagę podczas uczenia sieci, będzie on traktowany jako atrybut, z którym porównane zostaną grupy znalezione przez sieć. Wyłączenie zmiennej CLASS z procesu uczenia jest konieczne, aby późniejsza analiza porównawcza skupień znalezionych przez sieć oraz segmentów rynku z pola CLASS była miarodajna. Poza tym atrybutem, do analizy wybrano wszystkie pozostałe zmienne.

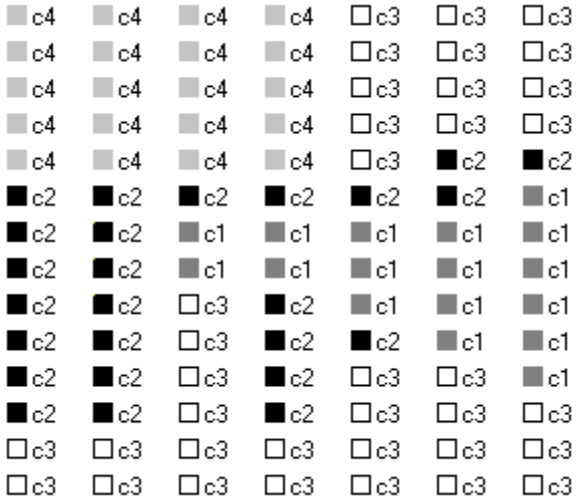
W czasie analizy utworzono kilkanaście sieci Kohonena, różniących się wielkością warstwy wyjściowej. Warstwa wejściowa w każdej z nich była taka sama i zawierała po jednym neuronie dla każdego atrybutu ilościowego i dwustanowego oraz trzy neurony dla atrybutu SERVICE (jeden neuron dla każdej wartości tego atrybutu).

Ostateczny rozmiar warstwy wyjściowej został ustalony eksperymentalnie. Minimalna liczba neuronów, jaka jest potrzebna do utworzenia mapy topologicznej, odzwierciedlającej cztery segmenty rynku, wynosi cztery (każdy neuron odpowiada innemu segmentowi). Jednak po przeprowadzeniu uczenia sieci o takiej liczbie neuronów wyjściowych rozmieszczonych w układzie 2×2 oraz po przeanalizowaniu sposobu działania nauczonej sieci, stwierdzono, że popełnia ona zbyt duży błąd w porównaniu z segmentami rynku w zmiennej CLASS. Każdy neuron był zwycięzcą dla przypadków należących do różnych segmentów rynku. Nie pozwalało to na precyzyjne przypisanie etykiet poszczególnych segmentów rynku do każdego z neuronów warstwy wyjściowej. Rozszerzenie rozmiaru mapy topologicznej ułatwiało zaobserwowanie skupisk rozpoznanych przez sieć oraz ich nazwanie. Ostatecznie do dalszych analiz wybrano sieć Kohonena, zawierającą 98 neuronów w warstwie wyjściowej, rozmieszczonych w układzie 14×7. Uczenie sieci Kohonena w Statistica Neural Networks przeprowadzono za pomocą algorytmu Kohonena. Uczenie składało się z dwóch etapów: uczenie wstępne i douczanie. Dla obu etapów eksperymentalnie ustalono wielkości trzech parametrów uczenia: liczbę epok, współczynnik uczenia, wielkość sąsiedztwa.

Po nauczeniu sieci Kohonena, uruchomiono ją dla wszystkich przypadków ze zbioru danych. Następnie nazwano każdy neuron warstwy wyjściowej, który był zwycięzcą dla co najmniej jednego przypadku ze zbioru danych. Podczas etykietowania neuronów korzystano z nazw segmentów rynku umieszczonych w zmiennej CLASS. Na tak powstałej mapie topologicznej można było łatwo zauważyć regiony, w których poszczególne neurony wygrywały dla przypadków należących tylko do jednego segmentu rynku. Jednak w niektórych obszarach mapy neurony okazały się być zwycięzcami dla przypadków należących do dwóch różnych segmentów rynku, co utrudniało to ustalenie granic pomiędzy skupiskami. Z tego względu, w następnym kroku przeanalizowano przypadki należące do trudno rozpoznawalnych obszarów mapy.

W czasie analizy porównawczej regionów mapy stwierdzono, że istnieją grupy wyrobów, które mają podobne cechy, a mimo to różne wartości atrybutu CLASS, czyli należą do różnych segmentów rynku. W związku z tym, mapę topologiczną poddano ponownemu etykie-

towaniu. Nowe etykiety odzwierciedlają granice skupisk zidentyfikowanych przez sieć Kohonena i biorą pod uwagę wyniki analiz porównawczych obszarów trudnych do interpretacji. Mapę topologiczną po ponownym etykietowaniu przedstawiono na rysunku 1.



Rys. 1. Mapa topologiczna sieci Kohonena po ponownym etykietowaniu neuronów
Fig. 1. Topological map of the Kohonen neural network after re-labeling of the neurons

Neurony po drugim etykietowaniu oznaczono etykietami „cn”, gdzie „n” jest numerem rozpoznanego segmentu rynku. Następnie każdemu przypadkowi ze zbioru danych przypisano nowy segment rynku na podstawie mapy topologicznej z rysunku 1. Informację o przynależności każdego wyrobu do jednego z nowych segmentów rynku zapisano w zbiorze danych, w postaci dodatkowego atrybutu o nazwie CLUSTER. Obok zmiennej CLASS, atrybut ten stanowi drugą zmienną zależną (zmienną celu). Wszystkie kolejne analizy, przeprowadzane za pomocą drzew decyzyjnych, zostały wykonane dla obu zmiennych zależnych. Pozwoliło to na ocenę wpływu sieci Kohonena na budowane drzewa decyzyjne.

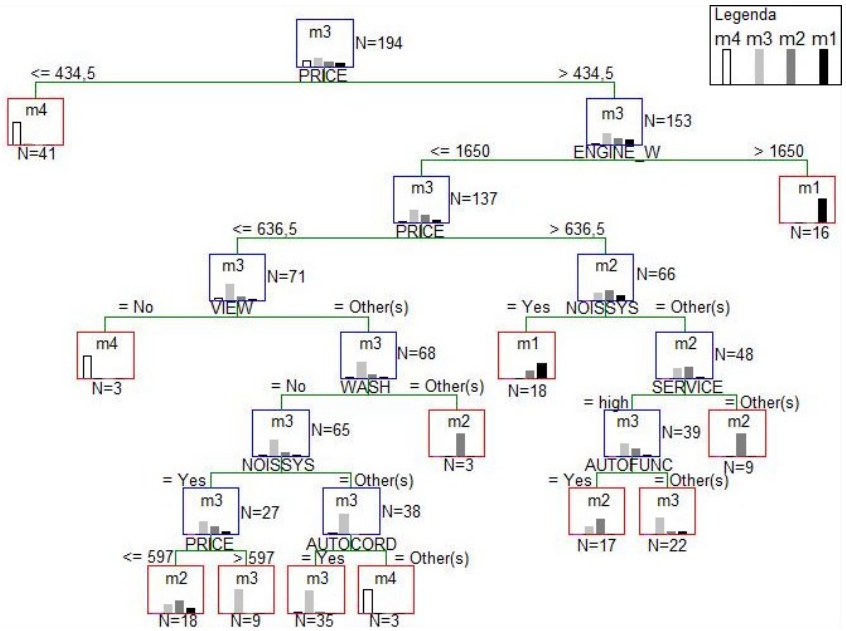
3.2. Segmentacja predykcyjna za pomocą drzew klasyfikacyjnych i regresyjnych CART

Segmentację predykcyjną rynku wykonano z wykorzystaniem pakietu programowego Statistica Data Miner, który zawiera duży nabór metod klasyfikacji i grupowania oraz budowy i implementacji odpowiednich modeli. Do realizacji zadań klasyfikacji w niniejszych badaniach najpierw zbudowano modele drzew klasyfikacyjnych i regresyjnych CART. Algorytmy pozwalające na wygenerowanie drzew decyzyjnych zalicza się do algorytmów uczenia z nauczycielem, tzw. uczenie nadzorowane. Wymaga to dostarczenia zbioru uczącego z jakościową zmienną celu, która dzieli przypadki ze zbioru uczącego na klasy. Algorytm uczy się, jakie wartości poszczególnych atrybutów (zmiennych wejściowych) odpowiadają każdej klasie zmiennej celu (zmiennej zależnej).

Drzewo decyzyjne jest grafem skierowanym, składającym się z węzłów decyzyjnych i liści, połączonych ze sobą za pomocą gałęzi. Każdy węzeł decyzyjny odpowiada atrybutowi ze zbioru danych. Z węzła wychodzą gałęzie symbolizujące możliwe wartości danego atrybutu. Węzły decyzyjne dokonują podziału zbioru danych na podzbiory, natomiast liście zawierają rozwiązania zadania klasyfikacyjnego, czyli przypisują podzbiorom przypadków konkretne klasy. CART jest algorytmem, który pozwala wygenerować drzewo ściśle binarne, gdzie każdy węzeł może dzielić zbiór danych tylko na dwa podzbiory. W eksperymencie porównano zdolność do predykcji drzew CART utworzonych dla zmiennych zależnych CLASS i CLUSTER.

Dla zmiennych zależnych CLASS i CLUSTER program wygenerował po 8 drzew decyzyjnych CART, o różnym stopniu przycięcia. Przycinanie drzewa decyzyjnego chroni go przed „przeuczeniem" (ang. *overfitting*), czyli nadmiernym dopasowaniem się do danych uczących. W celu wybrania najlepiej przyciętego drzewa wykorzystano V-krotny sprawdzian krzyżowy (ang. *V-fold cross validation*). Jako najlepiej przycięte drzewo zostaje wybrane to, które jest najmniej złożone i jednocześnie ma zbliżony do minimum koszt sprawdzianu krzyżowego. Według powyższego kryterium program wytypował modele przedstawione na rysunku 2 (dla zmiennej CLASS) oraz na rysunku 3 (dla zmiennej CLUSTER).

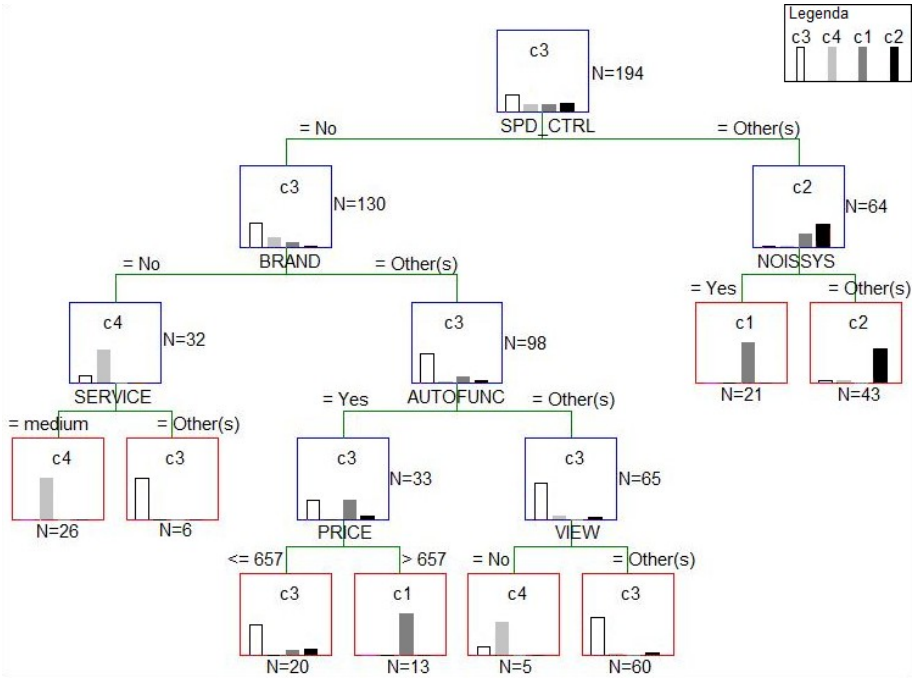
Histogramy na drzewie dla zmiennej zależnej CLUSTER pokazują, że przypadki klasyfikowane są z większą pewnością. Żaden z nich nie wskazuje, że do jednego liścia przydzielono zbliżoną ilość przypadków należących do różnych klas.



Rys. 2. Drzewo CART dla zmiennej zależnej CLASS
Fig. 2. CART tree for CLASS dependent variable

Pomimo tego, że algorytm CART korzystał podczas tworzenia obu drzew z tych samych predyktorów, to zmienne, które znalazły się w węzłach decyzyjnych, są inne. Atrybut najlepiej różnicujący zmienną CLASS to PRICE, natomiast dla zmiennej CLUSTER w korzeniu

drzewa znalazł się atrybut SPD_CTRL. Również to świadczy o innym podejściu do klasyfikacji wyrobów w przypadku obu drzew.



Rys. 3. Drzewo CART dla zmiennej zależnej CLUSTER
Fig. 3. CART tree for CLUSTER dependent variable

Porównując koszt sprawdzianu krzyżowego obu drzew oraz ich wielkości, można stwierdzić, że lepszą zdolność do predykcji ma drzewo utworzone dla zmiennej CLUSTER. Koszt sprawdzianu krzyżowego tego drzewa jest dwukrotnie mniejszy (wynosi 0,113) i jest ponaddwukrotnie mniejszy od kosztu dla zmiennej CLASS (0,289). Ponadto, model dla zmiennej CLUSTER jest mniej rozbudowany. Tak więc algorytm CART łatwiej znalazł zależności występujące pomiędzy atrybutami, analizując zbiór danych dla zmiennej zależnej CLUSTER.

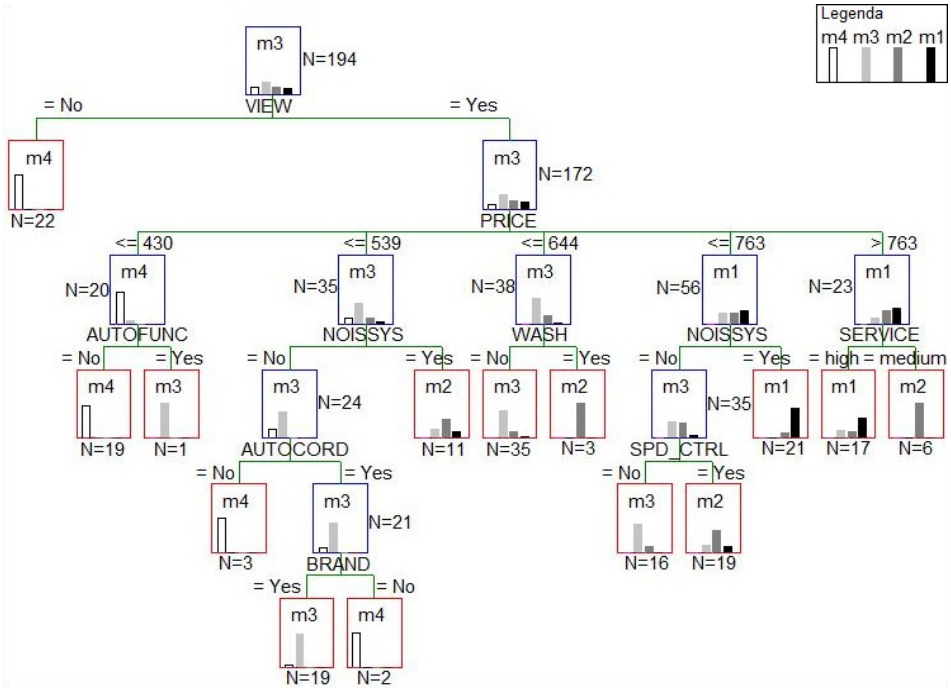
3.3. Segmentacja predykcyjna za pomocą drzew decyzyjnych CHAID

Następnie zadanie segmentacji rynków zostało wykonane z wykorzystaniem drzew decyzyjnych typu CHAID (ang. Chi-squared Automatic Interaction Detector). CHAID jest jednym z najstarszych algorytmów służących do tworzenia drzew decyzyjnych. Drzewo CHAID może służyć do zadań klasyfikacyjnych, do tworzenia drzew niebinarnych, czyli takich, w których z węzła decyzyjnego mogą wychodzić więcej niż dwie gałęzie. Jest to podstawowa cecha odróżniająca drzewa CHAID od drzew CART. Aby wybrać zmienne, które są umieszczone w węzłach decyzyjnych, algorytm posługuje się testem chi-kwadrat. Istotnymi parametrami podczas budowy drzewa CHAID są: minimalna liczba przypadków w węźle, który podlega podziałowi, prawdopodobieństwa podziału i łączenia kategorii oraz maksymalna liczba węzłów drzewa. Ich modyfikacja powoduje zmianę stopnia przycięcia drzewa. Wielkości tych parametrów wyznaczono eksperymentalnie, tak aby złożoność powstałych drzew była

zbliżona do utworzonego wcześniej drzewa CART dla zmiennej zależnej CLASS. Wygenerowane drzewo dla zmiennej CLASS przedstawiono na rysunku 4, natomiast model dla zmiennej CLUSTER jest pokazany na rysunku 5.

Analizując drzewo CHAID dla zmiennej CLASS, można zauważyć, że histogramy umieszczone w liściach drzewa ujawniają zależność, która wystąpiła już w drzewie CART dla tej zmiennej. Niektóre liście przydzielają do tej samej klasy kilka przypadków należących do różnych segmentów rynku. Taka tendencja zachowała się mimo tego, że oba drzewa różnią się od siebie rozmieszczeniem węzłów oraz doborem predyktorów w węzłach drzewa. Korzeń drzewa CHAID stanowi atrybut VIEW. Poza tym algorytm CHAID wykorzystał do budowy drzewa zmienne SPD_CTRL i BRAND, których brak jest w drzewie CART.

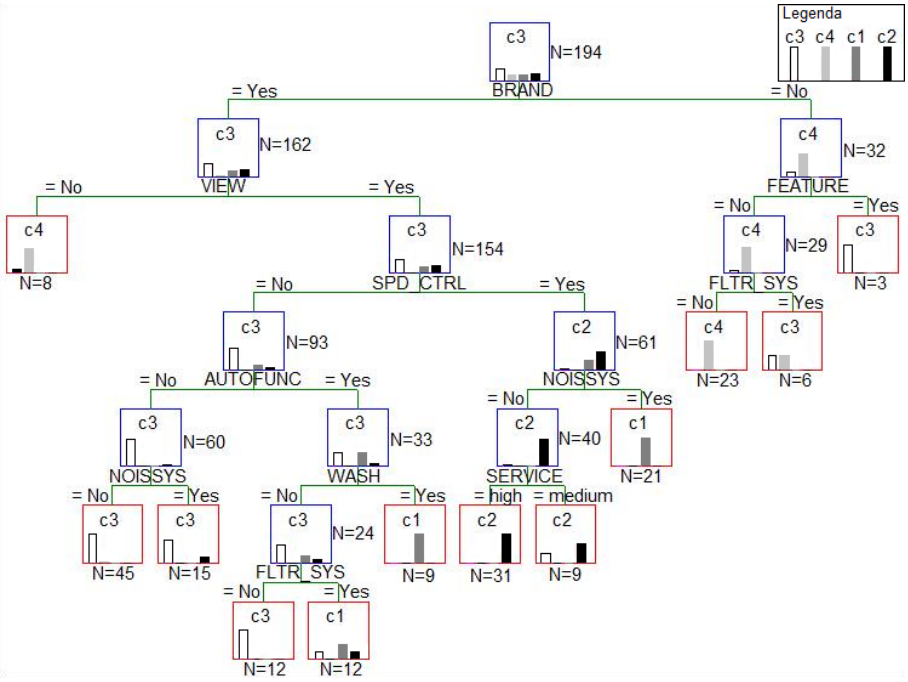
Porównując drzewa CHAID i CART dla zmiennej CLUSTER, można zauważyć kilka różnic. Przede wszystkim są to: wielkość drzew, dobór atrybutów w węzłach i występowanie w drzewie CHAID liści łączących przypadki należące do różnych klas. Liście te powodują większą niepewność podczas predykcji, odnotowaną już w zmiennej CLASS. Analizując ilość przypadków sklasyfikowanych do niewłaściwej klasy, można stwierdzić, że drzewo CART lepiej przystosowało się do danych uczących. Cechą wspólną obu drzew jest ich binarny charakter.



Rys. 4. Drzewo CHAID dla zmiennej zależnej CLASS
Fig. 4. CHAID tree for CLASS dependent variable

W przypadku drzew CHAID, V-krotny sprawdzian krzyżowy daje wartość liczbową określaną jako ryzyko estymacji. Miara ta zastępuję koszt sprawdzianu krzyżowego, używany do walidacji drzew CART. Program wyznaczył wartości ryzyka estymacji równe 0,244 dla modelu zmiennej CLASS oraz 0,144 w przypadku zmiennej CLUSTER. Tak więc przy

podobnej złożoności drzewa, algorytm CHAID łatwiej odnajduje zależności występujące w danych, używając zmiennej zależnej CLUSTER. Potwierdza to występowanie analogicznego zjawiska w przypadku drzew CART.



Rys. 5. Drzewo CHAID dla zmiennej zależnej CLUSTER
Fig. 5. CHAID tree for CLUSTER dependent variable

4. Podsumowanie

Do przeprowadzenia analiz drzew decyzyjnych wykorzystano dane opisujące zbiór wyrobów gospodarstwa domowego. Każdy przypadek w zbiorze danych należy do jednego z czterech segmentów rynku, opisanych zmienną zależną CLASS. Przed rozpoczęciem budowy modeli drzew decyzyjnych przeprowadzono grupowanie na zbiorze danych za pomocą sieci Kohonena. Grupowanie miało na celu zidentyfikowanie skupisk danych, odpowiadających segmentom rynku. Po znalezieniu skupisk przeanalizowano cechy produktów umieszczonych przez sieć w trudnych do zidentyfikowania obszarach mapy topologicznej. W wyniku analizy dla tych produktów, na nowo przypisano segmenty rynku. Informację o przynależności każdego wyrobu do jednego z nowych segmentów rynku zapisano w dodatkowej zmiennej zależnej CLUSTER, którą dołączono do analizowanego zbioru danych. Wszystkie kolejne analizy przeprowadzono równolegle dla wstępnie zdefiniowanych segmentów rynku w zmiennej CLASS oraz dla uporządkowanych przez sieć Kohonena segmentów rynku opisanych zmienną CLUSTER.

Do analizy wykorzystano dwa algorytmy używane do budowy drzew decyzyjnych: CART oraz CHAID. Algorytmy te zastosowano do budowy modeli predykcyjnych dla dwóch

zmiennych zależnych. Każdy z czterech otrzymanych modeli oceniono za pomocą miar oferowanych przez Statistica Data Miner. Wyniki tych analiz i porównania opracowanych modeli przedstawiono w poprzednich podrozdziałach. Wszystkie rodzaje drzew popełniały większy błąd predykcyjny podczas klasyfikacji wyrobów dla zmiennej zależnej CLASS. Klasyfikując przypadki ze względu na zmienną CLUSTER, błędy predykcji dla zbiorów testowych (lub dla V-krotnego sprawdzianu krzyżowego) były mniejsze.

Porównanie przedstawione w pracy ilustruje z jednej strony zależności pomiędzy dwoma różnymi algorytmami budującymi drzewa decyzyjne, a z drugiej, prezentuje różnicę pomiędzy zbiorem przypadków (dla odkurzaczy) sklasyfikowanych przez człowieka (zmienna CLASS) a tym samym zbiorem, który został uporządkowany za pomocą sieci Kohonena i wprowadzono do niego zmienną CLUSTER. Miało to na celu sprawdzenie, jaki będzie wpływ sieci Kohonena na wykorzystany zbiór danych uczących:

- jeśli modele drzew decyzyjnych dla obu zbiorów danych byłyby zbliżone, to wpływ sieci Kohonena jest znikomy,
- jeśli drzewa decyzyjne wygenerowane dla zbioru przypadków sklasyfikowanych przez sieć Kohonena popełniałyby większy błąd klasyfikacyjny niż drzewa decyzyjne dla zbioru oryginalnego, to wpływ sieci byłby negatywny (oznaczałoby to, że skupiska znalezione przez sieć Kohonena wyrażały mniej zależności pomiędzy przypadkami niż oryginalne klasy – segmenty rynku określone przez eksperta),
- jeśli natomiast błąd klasyfikacyjny drzew decyzyjnych dla zbioru klas zidentyfikowanych przez sieć Kohonena byłby mniejszy od błędu drzew decyzyjnych dla zbioru oryginalnego, wówczas mamy do czynienia z pozytywnym wpływem sieci w rozpatrywanym zbiorze danych (tak było w prezentowanej pracy).

Dalsze prace badawcze będą obejmować połączenie opracowanych modeli drzew decyzyjnych z innymi narzędziami sztucznej inteligencji, m.in. z systemami neuronowo-rozmytymi i sieciami neuronowymi. Połączenie tak odrębnych technik jest możliwe dzięki metodzie zwanej kontaminacją (ang. stacking). W przypadku badanego problemu możliwe jest przeprowadzenie odrębnej predykcji za pomocą drzew CART i CHAID oraz wzmacnianych drzew (boosted trees). Wyniki wygenerowane przez każdy z nich mogą zostać przekazane na wejście sieci neuronowej. Zadaniem sieci będzie nauczenie się, w jaki sposób połączyć wyniki poszczególnych drzew, tak aby ostateczny model predykcyjny dawał najlepsze rezultaty. Takie podejście nazywane jest „metauczeniem" (ang. *meta-learning*), ponieważ model zbiorczy uczy się na podstawie wyników nauczonych wcześniej modeli.

W niniejszej pracy zostały przedstawione wyniki wstępnych badań zastosowania narzędzi eksploracji danych do wspomagania segmentacji rynków, natomiast nie została przedstawiona ocena jakości wykonanych zadań grupowania i klasyfikacji, która będzie realizowana w przyszłości.

BIBLIOGRAFIA

1. Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa 2005.
2. Cios K., Pedrycz W., Świniarski R.: Data mining methods for knowledge Discovery. Kluwer, Norwell, MA 1998.
3. R. Bajek: Wykorzystanie metod eksploracji danych do budowy modeli scoringowych. Studia Informatica, Vol. 32, No. 2A (96), Gliwice 2011.
4. Kohonen T.: Self-organizing Maps. Proc. IEEE, 78, No. 9, 1990, s. 1464÷1480.
5. Kohonen T.: Self-organization and associative memory. Springer Verlag, Berlin 1989.
6. Larose D. T.: Odkrywanie wiedzy z danych. Wyd. Nauk. PWN, Warszawa 2006.
7. Li S.: The Development of a Hybrid Intelligent System for Developing Marketing Strategy. Decision Support Systems, Vol. 27, No. 4, 2000.
8. Migut G.: Zastosowanie technik analizy skupień i drzew decyzyjnych do segmentacji rynku. StatSoft, Materiały Seminarium StatSoft: Zastosowanie nowoczesnej analizy danych w marketingu i badaniach rynku, Kraków 2010.
9. Mynarski S.: Metody ilościowe i jakościowe badań rynkowych i marketingowych. StatSoft, Kraków 2010.
10. Nowak-Brzezińska A., Xięski T.: Grupowanie danych złożonych. Studia Informatica, Vol. 32, No. 2A (96), Gliwice 2011, s. 391÷401.
11. Setlak G.: The Fuzzy-Neuro Classifier for Decision Support. International Journal Information Theories and Applications, Pub. of the Institute of Information Theories and Applications FOI ITHEA, Vol. 15, No. 1, Sofia 2008, s. 22÷28.
12. Setlak G.: Intelligent Decision Support System. LOGOS, Kiev, 2004, (in Rus.), s. 250.
13. Stąpor K.: Metody klasyfikacji obiektów w wizji komputerowej. PWN, Warszawa 2011.
14. Stąpor K.. Automatyczna klasyfikacja obiektów. Wyd. Exit, Warszawa 2005.
15. Żurada J., Barski M., Jędruch W.: Sztuczne sieci neuronowe. PWN, Warszawa 1996, s. 375.

Wpłynęło do Redakcji 16 stycznia 2013 r.

Abstract

The main goal of this paper is to present and evaluate the possibility of using the methods and tools of artificial intelligence and data mining to analyze marketing data. The results of the analysis were needed to support decision-making in the process of market segmentation.

First and second chapter introduce theoretical foundation on market segmentation, data clustering, and data classification. The main part presents the results of market segmentation that can be used by the company producing household products. The data sets used in the study have been developed based on marketing research. The data includes vacuum cleaners' characteristics that are presented in table 1. Every product belongs to one of the four market segments that were determined during marketing research.

First part of the research was devoted to descriptive segmentation using Kohonen Self-Organizing Maps in Statistica Neural Networks. The task of the network was to carry out data clustering. The network has divided the set of vacuum cleaners onto four clusters, which are depicted in picture 1. The clusters are treated as the new market segments. After that, for each case from data set, the new market segment has been assigned.

Second part of the experiment covers predictive segmentation using the following types of classification trees: CART – Classification and Regression Tree and CHAID – Chi-squared Automatic Interaction Detector. These data mining techniques were used twice: for predictive the market segments obtained from marketing research, and for predictive the new market segments developed by Kohonen neural network. The analysis were carried out in Statistica Data Miner. The developed models were depicted in pictures 2-5.

The last part of the paper compared the results of solving classification problems using decision trees models. Finally further research plans have been described.

Adresy

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