Maciej Kutera*, Mirosława Lasek**

Zastosowanie metod analizy skupień w przeprowadzaniu segmentacji klientów na potrzeby kampanii reklamowych

Nadesłany: 6 lipca 2010 r. **Zaakceptowany:** 25 września 2010 r.

Streszczenie

Metody analizy skupień, obecnie zaliczane do metod eksploracji danych (ang. data mining), są coraz liczniejszą grupą metod, bardzo szybko rozwijającą się i znajdującą coraz więcej różnorodnych zastosowań. W artykule przedstawiamy badania ich przydatności w przeprowadzaniu segmentacji klientów na potrzeby prowadzenia kampanii reklamowych. Ukazujemy wyniki zastosowania do tego celu czterech wybranych metod, które ze względu na posiadane przez nie cechy wydawały się szczególnie obiecujące w działalności marketingowej. Jedna z analizowanych metod – metoda k-średnich z losowym wyborem początkowych środków skupień – dała nadspodziewanie przydatne wyniki segmentacji klientów dla prowadzenia kampanii reklamowych i te wyniki zostały w artykule przedstawione dość dokładnie. Inna z metod (hierarchiczna metoda średniej grupowej) okazała się mało użyteczna w przeprowadzaniu segmentacji klientów. Stąd podjęcie dalszych wysiłków, co do badań przydatności metod w przeprowadzaniu segmentacji klientów na potrzeby prowadzenia kampanii reklamowych uważamy za warte kontynuowania, szczególnie, że znalezienie rozwiązań w tym zakresie może przynieść konkretne i wymierne zyski w działaniach marketingowych.

Wprowadzenie

Wraz ze wzrostem mocy obliczeniowej komputerów i pojemności dysków twardych rosną bazy danych, a wraz z nimi ilość danych, które możemy zbierać i przetwarzać.

* Mgr, absolwent Wydziału Nauk Ekonomicznych Uniwersytetu Warszawskiego.
** Prof. dr hab., Wydział Nauk Ekonomicznych Uniwersytetu Warszawskiego.

W wielu przedsiębiorstwach gromadzone są ogromne ilości danych, bardzo często nie są one jednak w pełni wykorzystywane. Z pomocą może przyjść analiza skupień. Jest to jedna z metod eksploracji danych (ang. *data mining*), bardzo dynamicznie rozwijającej się dziedziny nauki w ostatnich latach.

Celem artykułu jest analiza przydatności, a także zalet i wad różnorodnych metod analizy skupień w segmentacji klientów na potrzeby prowadzenia kampanii reklamowych. Podjęta zostanie próba analizy i porównania także takich cech poszczególnych metod skupiania jak skuteczność w separowaniu różniących się skupień, wielkość skupień tworzonych za pomocą różnych metod.

Przedstawiane w artykule rozważania będą weryfikowane za pomocą analizy danych, pochodzących z badania *Diagnoza Społeczna 2007: Warunki i jakość życia Polaków* przeprowadzonego przez Radę Monitoringu Społecznego. Są to zarówno dane demograficzne, jak i deklaratywne, dotyczące zachowań, postaw i opinii osób uczestniczących w badaniu.

Zostanie podjęta próba podziału klientów, w oparciu o określone kryteria, na jednorodne grupy o podobnym popycie, aby skierować przekaz reklamowy odpowiednio do oczekiwań i potrzeb wydzielonych grup. Takie postępowanie może pomóc w podniesieniu konkurencyjności firmy oraz zwiększyć sprzedaż. Przy obecnej słabszej koniunkturze na światowych rynkach, może to mieć nawet kluczowe znaczenie dla utrzymania przedsiębiorstwa na rynku.

Kryteria segmentacji klientów można podzielić na „dotyczące konsumentów" oraz „dotyczące zakupywanych produktów".

Kryteria segmentacji klientów „dotyczące konsumentów" dzielone są na trzy grupy (Michalski, 2007):

- kryteria demograficzne, takie jak wiek, płeć, wielkość rodziny, narodowość,
- kryteria społeczno-ekonomiczne, takie jak dochód, zawód, wykształcenie, przynależność do określonej klasy społecznej,
- kryteria psychograficzne, dotyczące aktywności, sposobów spędzania wolnego czasu, zainteresowań, opinii.

W ostatnich latach coraz większe zainteresowanie wzbudza stosowanie kryteriów psychograficznych, tzw. „miękkich". Pozwalają one na pełniejsze zrozumienie klienta i sprostanie jego wymaganiom. W badaniach opisywanych w artykule wykorzystano zmienne wszystkich trzech wymienionych powyżej rodzajów.

Uwzględnione w badaniach kryteria „dotyczące zakupywanych produktów" dzielone są również na trzy grupy (Michalski, 2007):

- kryteria powiązane z wzorcami konsumpcji, obejmujące lojalność konsumentów wobec marki, częstotliwość korzystania z produktu, skłonność do innych produktów,
- kryteria dotyczące zakupu, dotyczące czasu zakupu, dostępności w punktach sprzedaży, wielkości jednorazowej partii zakupu, charakteru zakupu (impulsywny, codzienny, okazyjny),
- kryteria przedstawiające oferowane przez produkt korzyści, jak wiedza o produkcie, postrzeganie przez konsumenta zalety wynikające z posiadania produktu.

Przeprowadzenie segmentacji klientów powinno pomóc w podjęciu decyzji, do których z wyodrębnionych segmentów kierować przekaz reklamowy i w jakiej formie. W literaturze wyróżnia się trzy podstawowe sposoby kierowania przekazu reklamowego (Kotler, 2002):

- niezróżnicowana, gdzie ignorowane są różnice pomiędzy segmentami klientów i stosuje się jednolitą ofertę dla wszystkich klientów niezależnie od segmentu. Jest to uzasadnione rozwiązanie, gdy różnice pomiędzy segmentami są niewielkie;
- zróżnicowana, gdzie dokonuje się wyboru kilku spośród wyodrębnionych segmentów, a następnie tworzy oddzielną ofertę dla każdego z nich;
- skoncentrowana, gdzie uwaga zostaje skierowana na niewielką liczbę segmentów (często na jeden). Taka strategia jest stosowana najczęściej w przypadku, gdy do dyspozycji pozostają ściśle wyznaczone, ograniczone zasoby.

Wraz ze wzrostem zainteresowania technikami eksploracji danych pojawiło się wiele programów komputerowych wspomagających *data mining*, a w tym także analizę skupień. Na potrzeby przedstawianych w artykule analiz wykorzystano program SAS Enterprise Miner firmy SAS Institute Inc. z USA (Applied Analytics Using ..., 2008). Większość analiz przeprowadzono w SAS Enterprise Miner, natomiast niektóre w oparciu o własne procedury i makra napisane w języku programowania środowiska oprogramowania SAS, opracowanego przez SAS Institute Inc.

Wybrano oprogramowanie firmy SAS, ponieważ umożliwiło ono przetestowanie wszystkich badanych metod analizy skupień. Ponadto wzięto pod uwagę takie jego zalety, jak szybkość działania i dającą dobre rezultaty możliwość pracy z dużymi zbiorami danych. Uwzględniono także zaletę, polegająca na możliwości pisania własnych procedur, co pozwoliło dostosować algorytmy programu do potrzeb i specyfiki przeprowadzanych badań, a także przyspieszyć prowadzone prace.

1. Charakterystyka metod analizy skupień i ich przydatność do tworzenia segmentów o różnych własnościach

Analiza skupień, inaczej nazywana klasteryzacją lub po prostu grupowaniem (ang. *data clustering*, *cluster analysis*), to jedna z dziedzin statystyki, a zarazem podstawowe narzędzie analityczne służące do badań segmentacyjnych. Termin analiza skupień został po raz pierwszy użyty w 1939 r. przez R.C. Tryona (Tryon, 1970), natomiast szybki rozwój metod i zastosowań analizy skupień przypada na drugą poł. XX w. Klasteryzacja znalazła zastosowanie w wielu dziedzinach nauki, często bardzo odległych od siebie. Można wymienić takie dziedziny, jak: medycyna, psychologia, biologia, archeologia, geografia, genetyka, ekonomia. Analiza skupień znajduje zastosowanie wszędzie tam, gdzie posiadając duże zbiory obserwacji chcemy je podzielić na grupy przydatne dla określonych zastosowań. Segmentacja (grupowanie) klientów wydaje się dziedziną, w której analiza skupień może znaleźć zastosowanie przynoszące korzystne rezultaty.

Celem analizy skupień jest wykrycie wewnątrz zbioru obserwacji podzbiorów, zwanych także klastrami, w których obserwacje są do siebie podobne, różnią się natomiast od obserwacji zawartych w pozostałych klastrach. Analiza skupień pozwala wykrywać struk-

tury w danych, bez wyjaśnienia, dlaczego one występują. Klasteryzacja jest jedną z metod uczenia bez nadzoru („bez nauczyciela"). Uczenie bez nadzoru polega na tym, że zmienna zależna, której wartość jest wynikiem analizy, nie jest (i nie może być) bezpośrednio zaobserwowana. Celem przedstawianych w artykule badań jest wykrycie w zbiorze danych struktur lub skupień, których „nie widać" bez przeprowadzenia analizy. Dopiero po przeprowadzeniu klasteryzacji możemy zacząć opisywać i charakteryzować otrzymane skupienia (Internetowy Podręcznik Statystyki, 2009).

Metody analizy skupień są dzielone na metody hierarchiczne i podziałowe. Podstawowa różnica pomiędzy nimi wynika ze sposobu, w jaki są tworzone skupienia. W metodach hierarchicznych są to zagnieżdżone w sobie podziały, począwszy od całego zbioru do pojedynczych obserwacji, natomiast w metodach podziałowych wynikiem analizy jest tylko jeden podział o określonej liczbie klastrów (Jain, Murty, Flynn, 1999).

Liczba wszystkich algorytmów klasteryzacji sięga co najmniej kilkudziesięciu. Wiele z nich to algorytmy złożone i rozbudowane. Nie jest możliwe zastosowanie w badaniach i choćby pobieżne omówienie większej ich liczby w tym krótkim artykule. Ponadto opis metod jest dość obszernie przedstawiany w wielu pozycjach literatury i nie jest celem tego artykułu. Dla kompletności rozważań postanowiono przedstawić jedynie krótko metody stosowane na potrzeby analiz zamieszczonych dalej w artykule i prezentujące możliwie różnorodne podejścia do problemu analizy skupień, tj.:

- metody hierarchiczne,
- metodę k-średnich,
- samoorganizujące się mapy.

Metody hierarchiczne wykorzystują zagnieżdżone podziały. Metoda k-średnich daje jedne z najlepszych, najbardziej jednoznacznych wyników, jak stwierdzono według wcześniejszych badań (Jain, Murty, Flynn, 1999). Samoorganizujące się mapy są metodą wykorzystującą sieci neuronowe do tworzenia podziałów na skupienia.

Metody hierarchiczne były jednymi z pierwszych, dla których opracowano algorytmy tworzenia skupień (Jain, Dubes, 1988). Opracowano dwa rodzaje takich algorytmów: aglomeracyjne i deglomeracyjne. W algorytmach aglomeracyjnych zakłada się, że na początku tworzenia skupień, każda obserwacja jest oddzielnym, jednoelementowym klastrem. Następnie tworzona jest macierz odległości pomiędzy elementami zbioru i łączone są dwa znajdujące się najbliżej siebie. Proces jest powtarzany do momentu aż wszystkie obiekty zostaną połączone w jeden klaster. W algorytmach deglomeracyjnych postępowanie jest „odwrotne" niż w aglomeracyjnych. Zakładamy, że na początku tworzenia skupień obiektów stanowią one jeden klaster złożony ze wszystkich obiektów. Następnie w każdym kroku algorytmu wydzielamy jedną obserwację, najmniej pasująca do pozostałych. Tworzy ona oddzielne skupienie. Postępując w ten sposób dochodzimy do momentu, w którym każda obserwacja jest oddzielnym klastrem.

Graficzną ilustracją podziału na skupienia jest drzewo, nazywane dendrogramem. Na każdym poziomie drzewa znajdują się kolejne podziały, zaczynając od najbardziej szczegółowego, a kończąc na najbardziej ogólnym. Podział obserwacji na określoną liczbę skupień polega na „odcięciu" dendrogramu na danym poziomie. W ten sposób obserwacje zostają przyporządkowane do klastrów.

Zgodnie z powyżej przedstawionymi rozważaniami, posługując się metodami hierarchicznymi zbiór można podzielić na mniej lub więcej skupień w zależności od potrzeb i celu podziału na skupienia. Nie wymaga to powtórnej realizacji obliczeń, wystarczy „odciąć" dendrogram na innej wysokości. Utrudnienia pojawiają się wraz ze wzrostem obserwacji dzielonych na skupienia. Dendrogram staje się coraz bardziej złożony, trudny do analizy, nieczytelny, chociaż nadal można go wykorzystać do podziału zbioru obserwacji na skupienia przydatne do założonej analizy.

Istnienie wielu różnych algorytmów hierarchicznej analizy skupień wynika m.in. z możliwości stosowania różnych metod mierzenia odległości między skupieniami wieloelementowymi lub między skupieniem jednoelementowym i wieloelementowym.

Do stosowanych metod pomiaru odległości należą m.in. metody: najbliższego sąsiedztwa (ang. *single linkage*), najdalszego sąsiedztwa (ang. *complete linkage*), średniej grupowej (ang. *average linkage*), środka ciężkości (ang. *centroid method*), mediany (ang. *median method*), minimalnej wariancji Warda. Wybór miary odległości jest istotną decyzją, ponieważ zastosowana miara ma wpływ na charakter skupień, np. kształt tworzonych skupień: długie, rozciągnięte, czy też zwarte, kuliste. Poszczególne, wymienione tu metody pomiaru odległości są szczegółowo opisywane w licznych pozycjach literatury, nie będą więc tu omawiane. Z ich szczegółową charakterystyką, obejmującą stosowane sposoby mierzenia odległości (wraz ze wzorami pomiaru odległości), można zapoznać się np. w pracy (Kutera, 2010: 16–17). Przyjrzyjmy się natomiast skutkom zastosowania poszczególnych miar. Wybór metody najbliższego sąsiedztwa prowadzi zwykle do tworzenia długich, rozciągniętych skupień, co może dać efekt polegający na umieszczeniu zupełnie odmiennych obserwacji, w tych samych klastrach. Metoda najbliższego sąsiedztwa daje słabe wyniki w symulacjach Monte Carlo. W przypadku metody najdalszego sąsiedztwa, podobieństwo między skupieniami jest mierzone podobieństwem najmniej podobnych obserwacji z każdego klastra. Metoda ta ma skłonności do formowania zwartych, kulistych klastrów, o w przybliżeniu równych średnicach. Obserwacje nietypowe potrafią bardzo zaburzyć wyniki. Metoda średniej grupowej została opracowana, m.in. po to, aby zredukować zależność kryteriów sąsiedztwa od obserwacji odstających. Klastry wynikowe mają z reguły podobne rozproszenie wewnątrz skupień, a także łączone są ze sobą skupienia o małej wariancji. Metoda środka ciężkości jest bardziej odporna na obecność obserwacji odstających niż większość metod hierarchicznych. Jej wadą jest to, że podczas łączenia dwóch klastrów nierównej wielkości, mniejszy z nich staje się w znacznym stopniu zdominowany przez większy. Metoda mediany daje słabe wyniki podziału w symulacjach Monte Carlo. Skupienie powstałe z połączenia dwóch innych może być interpretowane jako pośrednie pomiędzy połączonymi. Metoda Warda ma skłonność do łączenia ze sobą klastrów o małej liczbie obserwacji i ma tendencję w kierunku tworzenia skupień z taką samą liczbą elementów. Jest bardzo wrażliwa na występowanie obserwacji nietypowych.

Pomimo prostoty stosowanych algorytmów, co jest niewątpliwą zaletą metod hierarchicznych, stwarzają one pewne trudności w stosowaniu. Nie ma jednego algorytmu, ani wskazówek, jaki algorytm wybrać, aby w przypadku danych o określonej charakterystyce otrzymać jak najlepsze rezultaty. Symulacje Monte Carlo wskazują na algorytmy: Warda, średniej grupowej i najdalszego sąsiedztwa jako na algorytmy najlepiej radzące sobie

z podziałami obserwacji na skupienia (Kutera, 2010: 18). Za istotną wadę metod hierarchicznych można uważać spadek efektywności wraz ze wzrostem liczby obserwacji. Metody te charakteryzują się kwadratową złożonością obliczeniową. Przez złożoność obliczeniową algorytmu rozumie się tu ilość zasobów komputera potrzebnych do rozwiązania problemu przez algorytm. Możemy mieć do czynienia ze złożonością pamięciową i czasową. Złożoność liniowa oznacza złożoność równą liczbie obserwacji zawartych w badaniu, natomiast złożoność kwadratowa jest równa kwadratowi liczby obserwacji. Wadą metod hierarchicznych jest fakt, że wraz ze wzrostem liczby obserwacji bardzo gwałtownie obniża się czytelność dendrogramu. Przy tysiącu obserwacji odczytanie podziałów jest bardzo trudne i pracochłonne, nie mówiąc o zbiorach liczących kilkadziesiąt lub więcej tysięcy obserwacji. Nie ma także możliwości korekty już utworzonych skupień w trakcie przeprowadzania badania, a w konsekwencji błędne przypisanie do skupienia nie może być skorygowane w kolejnym kroku algorytmu.

Następna z wymienionych metod analizy skupień, to metoda k-średnich. Jest ona przykładem metody kombinatorycznej. Podstawowy algorytm tej metody został opracowany przez J. B. MacQueena w 1967 roku (MacQueen, 1967). Algorytm składa się z czterech kroków postępowania:

- określana jest liczba grup – k, na którą ma być podzielony zbiór danych. Konieczność wyznaczenia a priori liczby skupień jest jedną z różnic między metodą k-średnich a uprzednio omówionymi metodami hierarchicznymi;
- wybierane są środki ciężkości dla tworzonych skupień. Jest to istotny element algorytmu, ponieważ wybór różnego rozmieszczenia środków ciężkości może dać odmienne wyniki metody. Jednym ze stosowanych sposobów jest wybór środków ciężkości możliwie jak najdalej oddalonych od siebie (Jain, Murty, Flynn, 1999). Innym sposobem jest wybór środków ciężkości otrzymanych za pomocą zastosowania metod hierarchicznych;
- dla każdej obserwacji ze zbioru danych znajdywany jest najbliżej położony środek ciężkości i obserwacja jest przypisana do skupienia z tym środkiem ciężkości;
- ponownie wyznaczane są środki ciężkości dla każdego z k skupień na podstawie należących do nich obserwacji i następuje powrót do kroku trzeciego algorytmu.

Kolejne kroki algorytmu są powtarzane aż do takiej sytuacji, gdy żadna z obserwacji nie zmieni już swojego skupienia.

Za jedną z podstawowych zalet metody k-średnich uważana jest jej wysoka wydajność (Kutera, 2010). Metoda charakteryzuje się liniową złożonością obliczeniową. Metoda jest efektywna nawet w przypadku bardzo dużych zbiorów danych, dając zazwyczaj w efekcie umożliwiające jednoznaczną interpretację rezultaty. Wadą metody jest konieczność podania przed jej zastosowaniem, liczby skupień, na jaką mają być podzielone obiekty. Obecnie opracowano wiele metod, które pomagają w wyborze liczby skupień przed rozpoczęciem grupowania. Przykładem może być metoda CCC, opracowana przez firmę SAS. Można także przeprowadzać symulacje, badając rezultaty stosowania różnej liczby skupień. Eksperymenty ze stosowaniem metody wskazują, że do jej słabszych stron można zaliczyć: dużą wrażliwość na wybór początkowych środków ciężkości, a także małą odporność na występowanie obserwacji odstających.

Samoorganizujące się mapy są rodzajem sieci neuronowych, które nadają się m.in. do tworzenia skupień obiektów. Algorytm tworzenia samoorganizujących się map został opracowany na początku lat 80. XX w. przez fińskiego profesora T. Kohonena (Kohonen, 1984; Kohonen, 2000). Stąd jest siecią neuronową zwaną często od nazwiska twórcy jej koncepcji siecią Kohonena. Najkrócej można ją scharakteryzować jako sieć samouczącą się z wbudowaną konkurencją i mechanizmem sąsiedztwa. Jest to sieć złożona z dwóch warstw neuronów: warstwy wejściowej i warstwy wyjściowej. Samouczenie polega na tym, że uczenie (trenowanie sieci) odbywa się w trybie „bez nauczyciela (ang. *unsupervised learning*; *self-organizing*), co oznacza, że dla podawanych danych wejściowych do treningu (tworzenia) sieci nie jest przedstawiana prawidłowa odpowiedź. Sieć nie jest zapoznawana z tym, jakie sygnały wyjściowe powinny odpowiadać wprowadzanym sygnałom wejściowym. Sieć tworzy się w toku odpowiednio opracowanego algorytmu samoorganizacji, stąd nazwa samoorganizujące się mapy, gdzie mapa jest wykresem, przedstawiającym wynik trenowania sieci (neurony warstwy wyjściowej). Tworzenie sieci odbywa się poprzez konkurencję między neuronami i modyfikację wag neuronów. Konkurencja jest mechanizmem powodującym, że neurony uczą się rozpoznawania sygnałów wejściowych i reagowania na sygnały wejściowe konkurując ze sobą. Neuron, który najsilniej zareaguje na dany sygnał wejściowy – im bardziej wagi neuronu są podobne do sygnałów wejściowych (wartości wejściowych), tym silniejsza reakcja – „zwycięża" w konkurencji rozpoznawania określonych sygnałów wejściowych. Inne neurony zostają „zwycięzcami" w rozpoznawaniu innych sygnałów (wartości) wejściowych. Sąsiedztwo jest tu rozumiane jako takie nauczanie sieci, że neurony sąsiadujące z neuronem „zwycięzcą" w rozpoznawaniu określonych sygnałów uczą się wraz z nim, chociaż mniej intensywnie. Takie trenowanie sieci powoduje, że sąsiadujące neurony będą reagowały na podobne sygnały (wartości) wejściowe. Zostaną utworzone skupienia (klastry), złożone z sąsiadujących neuronów dla sygnałów wejściowych o podobnej charakterystyce (Lasek, 2004).

Jak wynika z przedstawionego powyżej opisu, sieci Kohonena są uczone z wykorzystaniem algorytmu iteracyjnego. Rozpoczynając od wybranych losowo centrów radialnych, algorytm stopniowo modyfikuje je tak, aby odwzorować skupienia występujące w danych uczących. W pewnym stopniu jest to podobne do wyznaczania środków ciężkości w metodzie k-średnich, dlatego czasem sieci Kohonena są nazywane naturalnym uogólnieniem metody k-średnich (www.dataminer.pl/ textbook/stneunet.html). Tworzenie skupień polega więc na wyznaczeniu zwycięskiego neuronu i modyfikacji jego wektora wag w kierunku prezentowanego wektora wejściowego. Zwycięskim neuronem jest ten, którego wagi są najbliższe wektorowi danych wejściowych, ale istnieją dwa podstawowe sposoby modyfikacji neuronów wyjściowych. Jeden nosi nazwę „Zwycięzca Bierze Wszystko" (ang. *Winner Takes All*) i wówczas neuron najbardziej podobny do elementu prezentowanego zostaje przekształcony tak, aby jego wagi były jak najbardziej zbliżone do wektora wejściowego. Drugi sposób modyfikacji neuronów wyjściowych określany jest jako „Zwycięzca Bierze Większość" (ang. *Winner Takes Most*) i oznacza, że nie tylko neuron zwycięski, ale również jego otoczenie zostaje zmodyfikowane.

Zaletą samoorganizujących się map jest możliwość takiej reprezentacji graficznej skupień, która w przejrzysty i jasny sposób przedstawia na płaszczyźnie dwuwymiarowej przyporządkowanie obserwacji do klastrów. Jest nią często prezentowana w literaturze ilu-

stracja mapy topologicznej, będąca przedstawieniem warstwy wyjściowej neuronów radialnych (por. np. www.vias.org/tmdatanaleng/cc_ann_kohonen.html).

Ważnym zagadnieniem oprócz wyboru samej metody analizy skupień jest wybranie miary odległości (podobieństwa), która będzie określać, w jaki sposób będzie obliczane podobieństwo między dwoma elementami zbioru. Miara może mieć wpływ na kształt klastrów, ponieważ niektóre obiekty mogą znajdować się blisko siebie przy wykorzystaniu jednej miary, a daleko od siebie, jeżeli użyjemy innej miary. Z punktu widzenia algorytmu, wybór miary jest dowolny i tylko od przeprowadzającego badania, czy analizy zależy decyzja, jaka miara w danym przypadku będzie najodpowiedniejsza. Wzory obliczania stosowanych miar są powszechnie przedstawiane w literaturze i dlatego nie zostaną tu przytoczone. Można je znaleźć także w pracy (Kutera, 2010: 23–25). Często stosowaną miarą jest odległość euklidesowa. Sprawdza się dobrze w tych przypadkach, gdy w zbiorze danych możemy wyodrębnić zwarte i odosobnione klastry. Jej podstawową wadą jest dominacja zmiennej o największej skali. Rozwiązaniem tego problemu jest standaryzacja (normalizacja) zmiennych. Stosowany jest także kwadrat odległości euklidesowej. Różni się tym od odległości euklidesowej, że przypisuje większą wagę elementom bardziej oddalonym od siebie. Odległość euklidesowa oraz kwadrat odległości euklidesowej są chyba najpopularniejszymi i najczęściej stosowanymi miarami. Stosowanych jest jednak wiele innych miar. Zastosowanie miary, zwanej odległością miejską daje podobne wyniki jak w przypadku odległości euklidesowej. Zmniejszony jest jednak wpływ obserwacji odstających, ponieważ nie jest stosowana funkcja kwadratowa. Odległość Czebyszewa daje dobre wyniki i stosowana jest zazwyczaj wtedy, gdy chcemy oddzielić od siebie dwa elementy różniące się od siebie w jednym, dowolnym wymiarze. Odległość Minkowskiego jest uogólnieniem powyżej wymienionych miar poprzez wprowadzenie odpowiednich parametrów do wzoru służącego obliczeniu miary odległości. Stosujemy wybraną miarę ustalając parametry. Możemy przeprowadzać eksperymenty, badając zastosowanie różnych miar. Jeszcze inną miarą, stosowaną w analizie skupień jest kwadrat odległości Mahalanobisa. Jej zaletą jest to, że jest niezależna od skali zmiennych. Dodatkowym walorem metody jest to, że bierze pod uwagę korelacje w zbiorze danych. Gdy zmienne w zbiorze danych są dyskretne, stosowana jest miara niezgodności procentowej, zwana także od nazwisk jej autorów, miarą Sokala i Michenera. Wylicza, jaka część obserwacji jest do siebie niepodobna.

2. Przeprowadzanie segmentacji klientów za pomocą metod analizy skupień

W zapewnieniu dotarcia z przekazem reklamowym tak, aby skutecznie zainteresować właściwą grupę klientów, tj. potencjalnych nabywców produktów lub usług pomocne może być zastosowanie metod analizy skupień. Dzięki metodom analizy skupień będziemy mogli wyodrębnić grupy potencjalnych klientów tak, aby dla każdej z nich przygotować oddzielną ofertę, dopasowaną do danej grupy. Metody umożliwiają także wydzielenie osób, które prawdopodobnie nie będą wcale zainteresowane zakupem ani usługą. Bezpośrednim efektem skierowania odrębnych ofert do różnych grup potencjalnych klientów

jest obniżenie kosztów prowadzenia kampanii reklamowej, dzięki zwiększeniu szansy dotarcia do klienta, który będzie zainteresowany ofertą dopasowaną do jego zainteresowań.

Rozważania dotyczące przeprowadzania segmentacji klientów zostaną zilustrowane danymi, pochodzącymi z nadmienionego już wcześniej w artykule zbioru danych będących wynikiem badania *Diagnoza Społeczna 2007: Warunki i jakość życia Polaków* (Diagnoza Społeczna 2007: Warunki i jakość życia Polaków, 2009). Jest to cyklicznie przeprowadzane przez Radę Monitoringu Społecznego badanie ankietowe społeczeństwa polskiego. Jego celem jest uzupełnienie diagnozy opartej na wskaźnikach instytucjonalnych o kompleksowe dane na temat gospodarstw domowych oraz zachowań, postaw i kondycji psychicznej osób wchodzących w skład tych gospodarstw. Na potrzeby przeprowadzanych badań, dotyczących segmentacji klientów będzie wykorzystywany zbiór danych zawierający 183 zmienne i 18021 członków gospodarstw domowych. Zbiór zawiera zmienne o różnej skali pomiaru. Są to zarówno zmienne ciągłe (np. dochód, wiek), jak i dyskretne: porządkowe (np. poziom wykształcenia) oraz nominalne (np. województwo). W zbiorze danych znajdują się także odpowiedzi na pytania dotyczące postaw i zachowań społecznych respondentów.

Dla ilustracji przeprowadzania segmentacji klientów wybrano cztery metody. Eksperymenty z pozostałymi metodami planujemy przeprowadzić w dalszych badaniach. W pierwszej kolejności wybrano metody, które posiadają właściwości, sugerujące na szczególną przydatność w przeprowadzaniu segmentacji klientów na potrzeby prowadzenia kampanii reklamowych i wyniki ich zastosowania chcielibyśmy przedstawić w artykule. I tak spośród metod hierarchicznych wybrano metodę Warda oraz metodę średniej grupowej. Metoda minimalnej wariancji Warda została wybrana, ponieważ jest określana w literaturze jako najlepsza spośród metod hierarchicznych (Jain, Murty, Flynn, 1999). Daje również dobre wyniki w symulacjach Monte Carlo. Ponieważ okazała się skuteczna w licznych zastosowaniach, a zarazem biorąc pod uwagę jej cechy, można spodziewać się, że sprawdzi się także w badaniach opisywanych w artykule. Metoda średniej grupowej dobrze radzi sobie z obserwacjami odstającymi, co może mieć duże znaczenie w szczegółowej analizie, pozwalającej wyodrębnić specyficzne grupy klientów. Wykorzystano również metodę k-średnich, ponieważ jest to jedna z częściej stosowanych metod niehierarchicznych, dających możliwość dogłębnej analizy tworzonych skupień. Została tu zastosowana w dwóch wariantach. W pierwszym, początkowe środki skupień zostały wybrane losowo, natomiast w drugim są nimi środki skupień uzyskane w wyniku zastosowania metody Warda. Zastosowano także samoorganizujące się sieci neuronowe Kohonena jako przykład bardziej nowoczesnej, sprawdzającej się w wielu przypadkach metody w analizach skupień. Metoda ta jest warta próby zastosowania mimo swojej złożoności obliczeniowej, ponieważ tworzenie skupień przebiega w tym algorytmie na zupełnie odmiennej zasadzie niż w innych metodach analizy skupień i można sądzić, że może dać możliwość uzyskania dodatkowych, oryginalnych wyników, umożliwiających wnioskowanie dotyczące segmentacji klientów.

3. Przygotowanie wstępne danych do przeprowadzania segmentacji

Przed przystąpieniem do zastosowania algorytmów segmentacji klientów konieczne jest wstępne przygotowanie danych, obejmujące przedstawione poniżej podstawowe działania.

Należy dokonać wyboru zmiennych wykorzystywanych w tworzeniu skupień. Biorąc pod uwagę cel tworzenia skupień, czy też segmentów klientów, uwzględnia się zmienne demograficzne, zmienne dotyczące poziomu materialnego oraz zmienne dotyczące zwyczajów zakupowych badanych osób. W opisywanym w tym artykule badaniu zostały wybrane zmienne, tak jak je przedstawiono w zbiorze danych z badania *Diagnoza Społeczna 2007: Warunki i jakość życia Polaków* (Diagnoza Społeczna 2007: Warunki i jakość życia Polaków, 2009). Są to zmienne dotyczące:

- płci respondenta,
- znajomości języka angielskiego, mierzonej w trzystopniowej skali: zna język czynnie, zna język biernie, nie zna języka w ogóle,
- posiadania dostępu do Internetu,
- miesięcznego dochodu netto gospodarstwa respondenta w miesiącu poprzedzającym badanie,
- subiektywnej oceny poziomu materialnego respondenta w siedmiostopniowej skali: wspaniały, dobry, dosyć dobry, ani dobry ani zły, niezbyt dobry, zły, tragiczny,
- postaw respondenta, mierzonych w siedmiostopniowej skali: zdecydowanie tak, tak, raczej tak, ani tak ani nie, raczej nie, nie, zdecydowanie nie (uwzględniono tu trzy zmienne opisujące trzy postawy respondenta: lubię mieć rzeczy, których inni mogliby mi zazdrościć, lubię kupować rzeczy, które nie mają praktycznego znaczenia, samo robienie zakupów sprawia mi prawdziwą radość),
- zadowolenia respondenta z poziomu dostępnych dóbr i usług mierzonego w sześciostopniowej skali: bardzo zadowolony, zadowolony, dosyć zadowolony, dosyć niezadowolony, niezadowolony, bardzo niezadowolony,
- korzystania respondenta z komputera (używa lub nie),
- daty urodzenia respondenta,
- klasy miejscowości zamieszkania (w podziale na 6 kategorii: miasta powyżej 500 tys. mieszkańców, miasta o liczbie 200–500 tys. mieszkańców, miasta o liczbie 100–200 tys. mieszkańców, miasta o liczbie 20–100 tys. mieszkańców, miasta do 20 tys. mieszkańców, wieś),
- statusu społeczno-zawodowego respondenta (w podziale na 9 kategorii: pracownicy sektora publicznego, pracownicy sektora prywatnego, prywatni przedsiębiorcy, rolnicy, renciści, emeryci, uczniowie i studenci, bezrobotni, inni bierni zawodowo),
- poziomu wykształcenia (wyróżniono cztery kategorie: podstawowe i niższe, zasadnicze/gimnazjum, średnie, wyższe i policealne).

Aby usprawnić proces badania, a także poprawić czytelność otrzymanych wyników przydatne może być dokonanie przekształceń (transformacji) niektórych zmiennych. Przykładowo można utworzyć zmienną *wiek* ze zmiennej przedstawiającej *datę urodzenia*, połączyć zmienne: np. zmienną informującą, czy respondent ma Internet ze zmien-

ną, czy używa komputera (wówczas powstanie jedna zmienna binarna, gdzie odpowiedź „tak" oznaczałaby osoby mające dostęp do Internetu oraz korzystające z komputera, a „nie", przeciwny przypadek). Możemy usunąć obserwacje, reprezentujące te osoby, które nie udzieliły odpowiedzi na pytania dotyczące wymienionych wśród zmiennych *postaw respondentów*.

Skala odpowiedzi na niektóre pytania zawarte w badaniu może być uznana za zbyt szeroką. Różnica pomiędzy poszczególnymi odpowiedziami jest bardzo nieznaczna, a odpowiedź konkretnego respondenta może zależeć od jego sposobu wartościowania rzeczywistości, a niekoniecznie od faktycznego stanu rzeczy. Kolejnym krokiem może być więc scalenie niektórych odpowiedzi, tak aby uzyskać może mniej dokładne dane, ale bardziej skomasowane, lepiej oddające rzeczywistość. Przykłady możliwości scalania skal danych, które mogą dać korzyści polegające na jaśniejszej, mniej złożonej interpretacji, bez zbytniej utraty informacji przedstawiono w pracy (Kutera, 2010: 32–33).

Badanie korelacji pomiędzy zmiennymi, które chcemy wykorzystać do analizy skupień może stanowić podstawę usunięcia niektórych z nich, tak że nie będą wykorzystywane w segmentacji klientów, powodując niepotrzebną redundancję danych. Posiadane przez nas dane wskazują, że największy stopień korelacji (z liczącym się poziomem istotności) występuje pomiędzy znajomością języka angielskiego a wiekiem oraz pomiędzy znajomością języka angielskiego a korzystaniem z Internetu. Nie jest to zaskoczeniem, ponieważ to głównie młode osoby uczą się języków obcych, a językiem powszechnie wykorzystywanym w Internecie jest język angielski. Poziom skorelowania (Pearsona), nie jest jednak duży. Badanie korelacji nie dało podstaw do usuwania z analizy zmiennych, które zamierzamy wykorzystać w badaniach skupień klientów.

Przeprowadzanie analizy skupień za pomocą algorytmów grupowania wymaga rozwiązania problemu pojawiających się braków danych. Algorytmy te wymagają bowiem wykorzystywania zbiorów pozbawionych brakujących wartości, ponieważ w przypadku braków danych najczęściej automatycznie eliminują obserwacje z brakami danych.

W przypadku badań ankietowych, z jakimi mamy do czynienia w przedstawianej segmentacji klientów braki danych są powszechnym zjawiskiem. Najczęściej wynikają one z niechęci respondentów do udzielenia odpowiedzi na niektóre pytania ankiety. Ankiety najczęściej nie są kompletne. Przykładowo powszechnie zauważalny jest fakt, że ludzie nie są skłonni do ujawniania wysokości swoich zarobków i duże braki w danych mogą dotyczyć zmiennej opisującej dochód.

Najprostszym rozwiązaniem jest pominięcie obserwacji zawierających brakujące wartości. Musimy jednak pamiętać, że tak możemy tylko postąpić w przypadku posiadania odpowiednio licznego zbioru danych i stosunkowo niewielkiej liczby obserwacji z brakującymi danymi. Powszechnym rozwiązaniem problemu „brakujących danych" jest zastąpienie braków wartości zmiennej; w przypadku zmiennych ciągłych braki mogą być zastąpione np. jej wartością średnia, medianą, wartością środkową (max-min)/2), przyjętą stałą wartością; w przypadku zmiennych dyskretnych np. dominantą, stałą wartością, wylosowaną wartością zgodnie z rozkładem znanych wartości zmiennej. Opis licznych metod, także poza tu wymienionymi, ich rozbudowanymi wersjami i niejednokrotnie bardziej złożonych, mogących znaleźć zastosowanie do zastępowania brakujących wartości zmiennych można znaleźć w pracy (Applied Analytics Using…, 2008). Są to me-

tody zwane w środowisku SAS metodami imputacji danych i dostępnymi do stosowania w programie Enterprise Miner. W wielu przypadkach pomijanie obserwacji może być skutecznym sposobem, ponieważ sztuczne uzupełnianie wartości czasem grozi poważnym zaburzeniem wyników badania i może doprowadzić do nieprawdziwych wniosków. Metoda pomijania obserwacji traci swoją przydatność wraz ze wzrostem liczby obserwacji brakujących, ponieważ usuwamy coraz większy procent danych. Rozwiązaniem problemu brakujących wartości może być także usunięcie nie obserwacji, ale zrezygnowanie ze zmiennych, które mają dużą liczbę brakujących wartości. Metoda taka może jednak zaszkodzić naszym badaniom, jeżeli zrezygnujemy ze zmiennych mających duże znaczenie dla prowadzonej analizy.

W przypadku analizowanego przez nas zbioru danych, najwięcej brakujących obserwacji występuje w zmiennej dochodu netto – ponad 7% wszystkich obserwacji. Nie było to dla nas zaskakujące, ponieważ ludzie dosyć niechętnie podają swoje zarobki we wszelkiego rodzaju ankietach, czy badaniach, co jest znanym i potwierdzającym się powszechnie faktem w przeprowadzaniu wszelkich ankiet. W przypadku pozostałych zmiennych liczba obserwacji brakujących nie przekracza 1%. Jako metodę usunięcia braków danych wynikających ze zmiennej dochodu netto, nie byłoby rozsądne usunięcie tej zmiennej, ponieważ jest niewątpliwie zmienną o dużej wadze ze względu na cel badania. Po zbadaniu skutków usunięcia wszystkich obserwacji z brakami danych stwierdziliśmy, że struktura zbioru i rozkład jego podstawowych zmiennych nie uległ większej zmianie. Usunięcie obserwacji z brakami danych oznaczało utratę jedynie ponad 8% obserwacji. Jednocześnie średnie wartości zmiennych nie uległy większym zmianom. Z podanych względów w naszym przypadku pozbycie się brakujących wartości polegało na usunięciu obserwacji z brakami danych. Zbiór danych zawiera nadal ponad 11 000 obserwacji, co można uznać za wystarczające na potrzeby prowadzonych badań.

Na wynik przeprowadzenia skupiania obiektów duży wpływ może mieć występowanie obserwacji odstających (ang. *outliers*). Przykładowo efektem obecności obserwacji odstającej może być utworzenie jednoelementowego klastra, zawierającego tylko tę obserwację. W przypadku, gdy takich obserwacji jest więcej i wynikają z wartości różnych zmiennych, każda może zostać przypisana do oddzielnego skupienia, co już bardzo znacznie zaburzy wynik analizy skupień. Rozwiązaniem problemu obserwacji odstających jest ich usunięcie. Mogą być stosowane różne metody dla pozbycia się obserwacji odstających. Chyba najprostszą metodą jest usunięcie ustalonej liczby obserwacji, np. 5% o największych i najmniejszych wartościach. Wadą tej metody jest to, że nieświadomie możemy usunąć także obserwacje, których nie można traktować jako odstające. Inną metodą jest „ręczne", pojedyncze usuwanie wybieranych kolejno obserwacji, które budzą nasze wątpliwości. Jest to skuteczna metoda, jeżeli usuwanie obserwacji odbywa się rozważnie, ale może być bardzo pracochłonna w przypadku, gdy obserwacje opisywane są dużą liczbą zmiennych. Skuteczną metodą może być zbadanie odchylenia od średniej i usuwanie obserwacji odstających o przyjętą wielkość odchylenia standardowego, np. wielkości 3 odchyleń standardowych. Możliwą do stosowania metodą dla pozbycia się obserwacji odstających jest przeprowadzenie skupiania, bez eliminacji *outlierów*, sprawdzenie, czy występują skupienia jednoelementowe, usunięcie skupień jednoelementowych i dopiero przeprowadzenie przeznaczonej do stosowania segmentacji. Jeżeli usuwamy ob-

serwacje odstające, to musimy pamiętać, że zmieniamy strukturę używanego zbioru danych. Zazwyczaj nie są to duże zmiany. W przypadku przeprowadzanej przez nas segmentacji klientów na potrzeby prowadzenia kampanii reklamowych, celem analizy skupień jest pogrupowanie zbioru danych na duże i w miarę możliwości równoliczne skupienia osób, do których będziemy mogli kierować przekaz reklamowy. Tworzenie klastrów zawierających jedną lub kilka obserwacji nie wydaje się przydatne, poza chyba rzadkim przypadkiem „wyłowienia" pojedynczego klienta o specyficznych potrzebach.

W przypadku wykorzystywanego przez nas zbioru, zbadaliśmy występowanie obserwacji odstających analizując zmienną dochodu netto. Pozostałe zmienne, oprócz zmiennej, dotyczącej wieku respondenta, są zmiennymi dyskretnymi i analizy dotyczące obserwacji odstających ich nie dotyczą. Rozpiętość dochodu miesięcznego wynosi prawie 32000 PLN. Ponieważ przegląd danych wykazał, że wartości maksymalne bardzo szybko maleją postanowiliśmy usunąć 0,5% obserwacji najmniejszych oraz 0,5% obserwacji o największych wartościach zmiennej dochodu netto. Łącznie usunięto 114 obserwacji. Po takim usunięciu obserwacji minimalna wartość dochodu wynosi 420 zł, a maksymalna 10 600 zł (rozpiętość zmiennej teraz wynosi 10 180 zł).

Algorytmy przeprowadzania analizy skupień wymagają standaryzacji zmiennych wejściowych. Wszystkie użyte zmienne wejściowe powinny być zawarte w tym samym przedziale, np. [0, 1] lub [0, 100]. Pominięcie standaryzacji i wykorzystanie zmiennych mieszczących się w różnych przedziałach wartości spowoduje błędne przypisanie wag do poszczególnych zmiennych. Przykładowo, jeżeli wykorzystamy zmienną opisującą dochód i zmienną opisującą poziom wykształcenia, to rozpiętość dochodu będzie ponad tysiąc razy większa od rozpiętości poziomu wykształcenia. Takie też będą wagi zmiennych i w konsekwencji, zmienna opisująca poziom wykształcenia będzie bardzo mało istotna w porównaniu do dochodu. W naszej analizie zastosowaliśmy tzw. normalizację zmiennych, polegającą na odjęciu od każdej wartości zmiennej jej najniższej wartości, a następnie podzieleniu tej różnicy przez różnicę pomiędzy najwyższą a najniższą wartością. Wszystkie wartości dla każdej zmiennej po normalizacji, mieszczą się w przedziale [0, 1].

Zastosowanie metody analizy skupień wymaga wskazania liczby skupień (grup, czy tez segmentów), na jakie chcemy podzielić obiekty. Możemy dokonać wyboru liczby skupień subiektywnie (arbitralnie), opierając się na wiedzy, doświadczeniu i intuicji, biorąc pod uwagę cele, na jakie będziemy tworzyć skupienia. Spośród kryteriów mogących pomóc w wyborze liczby skupień, jako bardziej znane i częściej stosowane możemy wymienić:

- dendrogram – z wykresu dendrogramu można odczytać, na jakim poziomie powinniśmy wykonać „odcięcie", tzn. na ile skupień podzielić badany zbiór danych. Jest to metoda w dużym stopniu zależna od doświadczenia badacza;
- kryterium CCC (ang. *Cubic Clustering Criterion*) – według tego kryterium najodpowiedniejsza liczba skupień jest wyznaczona poprzez lokalne, dodatnie maksimum;
- statystyka pseudo-F – najlepiej dopasowana liczba skupień jest wyznaczona poprzez lokalne maksimum;

- test pseudo-T^2 – najlepiej dopasowana liczba skupień jest wyznaczona poprzez lokalne minimum.

Powyżej przedstawione kryteria nie dostarczają dokładnej odpowiedzi, na jaką liczbę skupień powinniśmy podzielić obiekty. Mogą tylko pomóc przy wyborze właściwej liczby skupień. Bardzo często zdarza się, ze różne kryteria wskazują na różną liczbę skupień. Dokonując wyboru liczby skupień podziału obiektów należy pamiętać o celu badania. Przykładowo, podział na dziewięć skupień bazy klientów małego sklepu internetowego może okazać się zbyt szczegółowy, ponieważ zarządzanie tyloma segmentami klientów będzie zbyt czasochłonne i kosztowne. Przy podejmowaniu decyzji o liczbie skupień ważne jest doświadczenie i zdroworozsądkowe podejście. Dla naszych danych i zastosowania hierarchicznej metody średniej grupowej, sporządzony dendrogram sugeruje od czterech do siedmiu klastrów jako optymalną ich liczbę. Pozostałe z wymienionych kryteriów, sugerują następujące liczby skupień:

- 4 lub 7 skupień w przypadku kryterium CCC,
- 4 lub 7 skupień w przypadku statystyki pseudo-F (a więc tak samo jak kryterium CCC),
- 4 lub 6 skupień w przypadku testu pseudo-T^2.

Rozważając wszystkie rozpatrywane kryteria jesteśmy skłonni podjąć decyzję o wyborze liczby skupień: 4 lub 7. Biorąc pod uwagę cel naszej analizy wybór czterech klastrów wydaje się być bardziej przydatny. Zbyt duże rozdrobnienie bazy klientów mogłoby stworzyć znaczne utrudnienie analizy, podnosząc jednocześnie koszty dotarcia do klientów z odpowiednim przekazem reklamowym.

4. Interpretacja wyników segmentacji klientów na potrzeby prowadzenia kampanii reklamowej

Wyniki segmentacji klientów przeprowadzone za pomocą dwóch metod hierarchicznych (jednej z wykorzystaniem metody średniej grupowej oraz drugiej z wykorzystaniem metody minimalnej wariancji Warda) oraz dwóch metod podziałowych: metody k-średnich oraz samoorganizujących się map Kohonena zostały wraz z charakterystyką otrzymywanych skupień szczegółowo przedyskutowane w pracy (Kutera, 2010).

Za pomocą hierarchicznej metody średniej grupowej przeprowadziliśmy podział obserwacji (klientów) na cztery klastry. Podział na przyjęte przez nas cztery klastry okazał się bardzo nierówny pod względem liczby przydzielonych klientów do poszczególnych klastrów. Największy klaster zawierał blisko 70% wszystkich klientów, najmniejszy składał się z 18 osób. Pomimo bardzo starannego przygotowania danych nie uzyskaliśmy zbyt dobrego wyniku pod względem formalnym. Jednakże mając na uwadze cel naszej analizy, jakim jest podział klientów na grupy tak, że do każdej grupy kierujemy oddzielny przekaz marketingowy, uzyskane wyniki mogą być jednak przydatne. Uzyskane grupy różnią się między sobą dość znacznie. Szczegółowa analiza wartości zmiennych osób z poszczególnych grup wskazała, że ofertę warto skierować do grup, oznaczonych umownie jako grupa 2, 3 i 4. Grupa 1, pomimo, że jest to właśnie najliczniejsza grupa, nie wydaje się być grupą, do której celowe byłoby przesłanie oferty reklamowej. Znajdują się

w niej bowiem osoby o najniższych dochodach, starsze, mało podatne na reklamy i stąd mało perspektywiczne jako klienci.

Kolejną zastosowaną metodą była metoda hierarchiczna z zastosowaniem algorytmu minimalnej wariancji Warda. Po przeanalizowaniu wyników kryteriów wyboru liczby skupień, zdecydowaliśmy podobnie jak w poprzednio opisanym przypadku metody hierarchicznej średniej grupowej, podział zbioru obserwacji na cztery skupienia. Taki wybór wynikał z analizy wyników kryteriów wyboru liczby skupień, ale ma jeszcze dodatkowo tę zaletę, że z pewnością pomoże porównać wyniki uzyskane za pomocą metody Warda z poprzednio przedstawioną metodą średniej grupowej.

Wyniki uzyskane za pomocą metody Warda znacznie różnią się od uzyskanych uprzednio wyników za pomocą średniej grupowej. Otrzymane skupienia klientów mają tym razem zbliżoną liczbę elementów. Analiza uzyskanych skupień wskazuje, że wyników podziału na skupienia nie można jednak uznać za w pełni satysfakcjonujące. Podział na skupienia jest bowiem w tym przypadku oparty przede wszystkim na zróżnicowaniu wartości, jakie przyjmują tylko trzy spośród wszystkich wytypowanych do analizy zmiennych. Te trzy zmienne, to zmienna określająca płeć, określająca fakt, czy respondent wykorzystuje Internet oraz zmienna opisująca zadowolenie z poziomu dostępnych dóbr i usług. Na podstawie otrzymanych wyników można sądzić, że dwie spośród czterech wydzielonych grup klientów można rozważać jako grupy, do których warto skierować przekaz reklamowy. Wydaje się, że pozostałe dwie grupy można pominąć w prowadzonej kampanii reklamowej, ponieważ osoby znajdujące się w nich prawdopodobnie nie będą zainteresowane żadną ofertą.

Jako kolejna, została wykorzystana metoda k-średnich. Zastosowano tę metodę wybierając dwa nieco różniące się, realizujące ją algorytmy. W przypadku pierwszego algorytmu, początkowe środki skupień są wybierane losowo. W przypadku drugiego algorytmu, środki skupień są najpierw wybierane za pomocą metody Warda, a w następnej kolejności jest przeprowadzana klasteryzacja za pomocą metody k-średnich.

Posługując się metodą k-średnich, która nie jest metodą hierarchicznego skupiania, wybierając liczbę skupień nie możemy wykorzystać dendrogramu. Możemy posłużyć się kryterium CCC oraz statystyką pseudo-F. Zarówno w przypadku kryterium CCC, jak i statystyki pseudo-F zostało zasugerowanych pięć lub osiem skupień. Wybraliśmy mniejszą liczbę skupień (pięć), co zapewne nie tylko ułatwi wskazanie segmentów, do których powinniśmy skierować przekaz reklamowy, ale także porównanie wyników z wynikami otrzymanymi za pomocą metod hierarchicznych (przypomnijmy, że w przypadku metod hierarchicznych został w obu rozważonych przypadkach wybrany podział na cztery segmenty klientów).

Z analizy wyników zastosowania metody k-średnich wynika, że spośród pięciu rozważanych skupień, charakterystyka czterech wskazuje na możliwość uzyskania korzystnych efektów w przypadku skierowania do osób z tych czterech skupień odpowiedniego przekazu reklamowego. Stosując metodę k-średnich otrzymaliśmy klastry o zbliżonej liczbie obserwacji – klientów, w każdym klastrze. Nie ma zmiennych, których wartości są bardzo mocno zróżnicowane pomiędzy klastrami, natomiast można zauważyć, że albo dana wartość zmiennej występuje w klastrze, albo nie występuje. Biorąc pod uwagę

przedstawione fakty, otrzymany podział uznaliśmy za mogący być bardzo przydatnym w prowadzeniu kampanii reklamowych.

Przeprowadziliśmy teraz analizę, stosując drugi z wymienionych algorytmów metody k-średnich, a więc algorytm, gdy początkowe środki skupień nie są wybierane losowo, ale za pomocą realizacji hierarchicznej metody Warda i uzyskane w ten sposób służą za startowe środki skupień w metodzie k-średnich. Wybraliśmy możliwość podziału klientów na pięć skupień, aby porównać wyniki zastosowania metody k-średnich, gdy zostały zastosowane różne algorytmy wyboru początkowych środków skupień.

Po zmianie początkowych środków ciężkości, zdecydowanie zmieniła się charakterystyka otrzymanych klastrów. Charakterystyka trzech z nich wskazuje, że skierowanie przekazu reklamowego do osób należących do tych klastrów może dać pozytywne wyniki. Dwa klastry, zawierają osoby o charakterystyce wskazującej, że kierowanie w tym przypadku przekazu reklamowego nie jest celowe. Liczba obserwacji w poszczególnych klastrach jest do siebie zbliżona. Podobnie jak w przypadku zastosowania hierarchicznej metody Warda pojawiły się zmienne, których wartości różnią się bardzo silnie pomiędzy klastrami. Są to zmienne dotyczące: płci respondenta, używania Internetu i zadowolenia respondenta z poziomu dostępnych dóbr i usług. Nielosowy wybór początkowych środków skupień nie dał lepszej segmentacji klientów na potrzeby kierowania przekazu reklamowego do klientów.

Ostatnią zastosowaną przez nas metodą segmentacji klientów była metoda samoorganizującej się mapy Kohonena. Spośród dostępnych algorytmów, został zastosowany algorytm grupujący *Kohonen Self-Organizing Map*. W przypadku samoorganizujących się map przed rozpoczęciem tworzenia segmentów, algorytm wymaga podania liczby wierszy i kolumn (wymiarów warstwy wyjściowej sieci). Przemnożenie liczby wierszy i kolumn wskazuje na liczbę tworzonych segmentów. Dla potrzeb naszego badania ustaliliśmy liczbę wierszy na dwa oraz liczbę kolumn także na dwa, a więc w rezultacie otrzymamy cztery segmenty, czy też skupienia (grupy) klientów. Wybór czterech segmentów ułatwi porównanie wyników grupowania za pomocą sieci Kohonena z wynikami otrzymanymi za pomocą innych metod. W wyniku zastosowania sieci Kohonena otrzymaliśmy klastry, które można uznać za znacznie od siebie oddalone. Uzyskane grupy mają zbliżoną liczbę elementów. Nie ma klastrów z bardzo małą lub bardzo dużą liczbą obserwacji (osób, czy potencjalnych klientów). Podobnie, jak w przypadku metody k-średnich z nielosowym wyborem środków skupień, podział na skupienia jest także oparty w dużym stopniu na zróżnicowaniu wartości zmiennych określających płeć, korzystanie z Internetu, zadowolenie respondenta z poziomu dostępnych dóbr i usług. Szczegółowa analiza wyników doprowadza do wniosku, że uzyskaliśmy dwa skupienia osób, do których nie wydaje się celowe skierowanie przekazu reklamowego i dwa skupienia osób, do których skierowanie przekazu reklamowego wydaje się celowe i może przynieść dobre efekty.

Przegląd wyników uzyskanych za pomocą różnych metod analizy skupień wskazuje, że najmniej przydatne na potrzeby prowadzenia kampanii reklamowych wydają się wyniki uzyskane za pomocą metody średniej grupowej. Skupienia, pod względem liczebności, zostały wydzielone bardzo nierównomiernie, największy klaster zawiera 70% obserwacji, a najmniejszy niecałe 0,5% (tylko 18 obserwacji). Najliczniejszy klaster okazał się grupą osób, do której nie wydaje się celowe kierowanie przekazu reklamowego. Po-

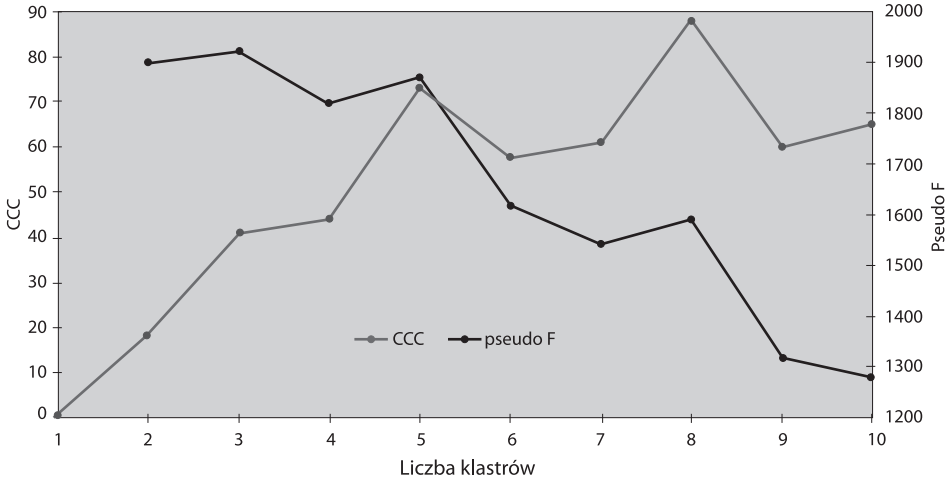
mimo bardzo starannego przygotowania danych przed rozpoczęciem analizy skupień nie udało się osiągnąć przydatnej segmentacji na potrzeby organizacji kampanii reklamowych za pomocą metody średniej grupowej.

Wyniki uzyskane za pomocą pozostałych metod można uznać za zbliżone do siebie. Najbardziej podobne pod względem liczebności poszczególnych skupień są wyniki uzyskane w metodzie Warda i za pomocą sieci Kohonena. Należy zauważyć, że żadnych wyników segmentacji, oprócz uzyskanej za pomocą metody średniej grupowej, nie można uznać jako całkowicie nieprzydatnych w prowadzeniu kampanii reklamowych. Uzyskane wyniki za pomocą poszczególnych metod, dostarczają charakterystyk różnych grup klientów i w przypadku prowadzenia konkretnej kampanii reklamowej mogą dostarczyć cennych sugestii, co do kierowania nieraz bardzo specyficznych przekazów reklamowych do specyficznych grup osób, wyłonionych, za pomocą akurat jednej, konkretnej metody analizy skupień, a nie wyodrębnionych przez inne dostępne metody.

Analiza wyników uzyskanych za pomocą różnych metod skupiania wydaje się wskazywać na metodę k-średnich z losowym wyborem początkowych środków skupień jako na metodę, która dała bardziej szczegółowe i konkretne wskazówki kierowania przekazów reklamowych do wydzielonych za jej pomocą grup osób. Otrzymaliśmy cztery różne grupy osób, do których warto skierować przekaz reklamowy i jedną grupę, do której skierowanie przekazu reklamowego raczej nie da efektów. Przedstawmy bardziej szczegółowo wyniki tej metody jako metody, która dała wyniki dość precyzyjnego podziału osób na skupienia, do których można kierować odpowiednio uprofilowane przekazy reklamowe i stąd wydające się na przydatniejsze od wyników innych metod.

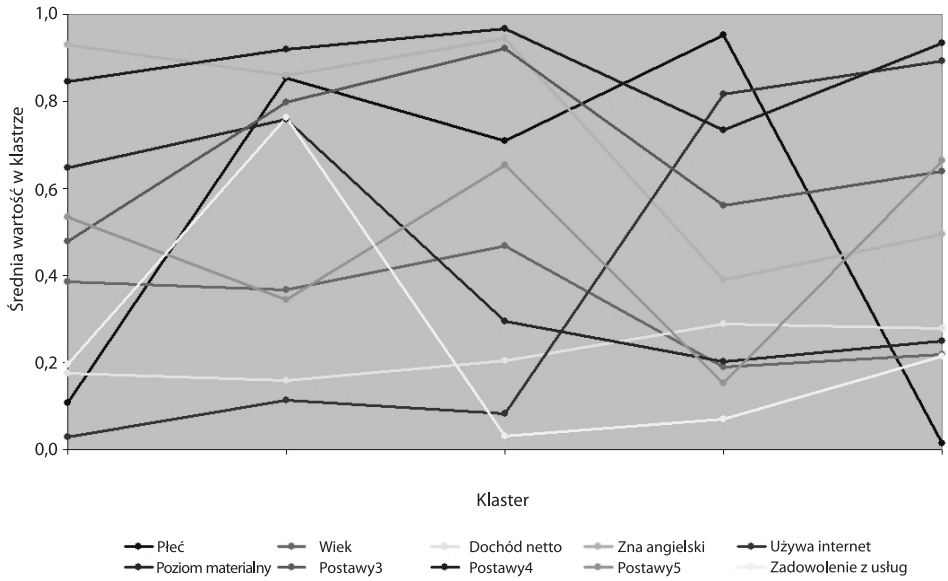
Rys. 1 ilustruje zastosowanie kryterium statystyki pseudo-F oraz kryterium CCC w wyborze liczby skupień. Widzimy, że sugerowany jest wybór pięciu lub ośmiu skupień. Jak już wspomniano w artykule, ze względu na wystarczającą liczbę skupień dla roz-

Rys. 1. Kryteria wyboru liczby skupień – metoda k-średnich



Źródło: opracowanie własne.

Rys. 2. Zróżnicowanie zmiennych – metoda k-średnich z losowym wyborem środków skupień (postawy oznaczają: Postawy3 – *lubię mieć rzeczy, których inni mogliby mi zazdrościć*; Postawy4 – *lubię kupować rzeczy, które nie mają praktycznego znaczenia*; Postawy5 – *samo robienie zakupów sprawia mi prawdziwą*



Źródło: opracowanie własne.

ważenia efektywnej akcji reklamowej oraz zamiar porównywania wyników tej metody z wynikami innych metod, wybraliśmy mniejszą liczbę pięciu skupień.

Na rys. 2 przedstawiono wykres średnich wartości zmiennych w poszczególnych klastrach. Widoczne są wyraźne różnice wartości zmiennych w skupieniach.

Możemy przeanalizować poszczególne, uzyskane skupienia. Ich szczegółową charakterystykę zamieszczono w tabeli 1.

Klaster 1 składa się z 2255 obserwacji. Wśród znajdujących się w nim osób 89% stanowią mężczyźni, a 11% kobiety. W opisywanej grupie przeważają mieszkańcy wsi, osoby z wykształceniem zasadniczym i niższym, renciści i bezrobotni. Ze szczegółowej charakterystyki klastra 1 wynika, że należą do niego w większości starsi mężczyźni, słabo wykształceni, o niskich dochodach. Większość osób określa swój status materialny jako zły. Analizując ten klaster warto zauważyć, że znaczna część osób, bo prawie połowa całej grupy, lubi mieć rzeczy, których inni mogliby im zazdrościć. Wydaje się, że sporządzając dla osób z tego klastra ofertę reklamową można ten fakt wykorzystać, umieszczają w ofercie odpowiednie przedmioty.

W drugim klastrze znalazło się 1942 osób. Aż 85% osób stanowią kobiety (a więc zaledwie 15% mężczyźni). W opisywanym tu, drugim klastrze, podobnie jak w klastrze pierwszym przeważają osoby ze wsi, o wykształceniu zasadniczym i niższym, rolnicy, renciści i bezrobotni. Z przedstawionego powyżej opisu obu klastrów nietrudno zauważyć podobieństwa między nimi. Podstawową różnicą jest to, że osoby należące do klastra

2, to kobiety, podczas gdy osobami należącymi do klastra 1 byli głównie mężczyźni. Kobiety zaliczone do klastra 2 są bardziej niezadowolone ze swojego poziomu materialnego niż mężczyźni z klastra 1. Pomimo tego, dla ponad połowy osób klastra 2, samo robienie zakupów jest przyjemnością. Dominują w tym klastrze starsze kobiety. Pomimo, że ponad połowa osób lubi robić zakupy, to tylko niecałe 5% lubi kupować rzeczy niepraktyczne. Można wnioskować, że jest to grupa, do której celowe jest skierowanie oferty reklamowej produktów tanich oraz produktów codziennego użytku, artykułów gospodarstwa domowego, ubrań (z powodu niskich dochodów powinny to być towary z tzw. „dolnej półki".

Do klastra 3 zostało zaliczonych 3931 obserwacji. Wśród osób tej grupy, 71% stanowią kobiety, a 29% – mężczyźni. Wśród osób zaliczanych do opisywanego klastra przeważają osoby z wykształceniem podstawowym, emeryci i renciści. Opisywany klaster (trzeci) stanowi grupę osób starszych, o średnich dochodach, w większości nie korzystających z Internetu i nie znających języka angielskiego. Respondenci tego klastra deklarują, że nie lubią kupować zbędnych rzeczy. Robienie zakupów nie sprawia im przyjemności. Są to przede wszystkim emeryci i renciści, nie lubiący robić zakupów. Jednocześnie wskazują na swój poziom materialny jako na dobry i na zadowolenie z poziomu oferowanych im dóbr i usług. Prawdopodobne jest, że osoby z opisywanej grupy nie będą zainteresowane otrzymywaniem informacji reklamowych i będą podchodzić z nieufnością do przekazów reklamowych.

Klaster 4 tworzy 1603 obserwacje. W zdecydowanej większości grupę tworzą kobiety (95% osób), mężczyzn w tej grupie jest zaledwie 5%. Osoby opisywanego tu 4. klastra, to przede wszystkim mieszkańcy miast liczących powyżej 20 tys. osób. Są to osoby z wykształceniem średnim i wyższym, uczniowie, studenci, pracownicy sektora publicznego. Przedstawiając ogólną charakterystykę 4. klastra można stwierdzić, że należą do niego w przeważającej większości młode, dobrze wykształcone kobiety z miast, o dość wysokich dochodach. Znają język angielski. Korzystają z Internetu. Lubią robić zakupy, kupować rzeczy niepotrzebne („niepraktyczne"), a także takie, których inni mogą im zazdrościć. Można wnioskować, że jest to grupa osób podatna na reklamę, do której warto kierować przekazy reklamowe. Na podstawie charakterystyki osób z tej grupy, wydaje się zasadny wniosek, że aby zachęcić te osoby do zakupu dóbr lub korzystania z usług, reklama powinna być oryginalna, niestandardowa, zawierająca ofertę ubrań, kosmetyków, lubianych przez kobiety „drobiazgów". Z charakterystyki osób w grupie można z dużym prawdopodobieństwem przewidzieć, że z zainteresowaniem spotkają się artykuły markowe i modne.

Klaster 5 składa się z 1631 obserwacji (osób). W zdecydowanej większości są to mężczyźni (98% osób). Tylko 2% osób stanowią kobiety. Klaster 5 to grupa osób będących mieszkańcami miast powyżej 20 tys., z wykształceniem średnim i wyższym, prywatni przedsiębiorcy, studenci, pracownicy sektora publicznego i prywatnego. Pod wieloma względami grupa ta jest podobna do poprzedniej. Podstawowa różnica, to fakt, że jest prawie w całości złożona z mężczyzn, podczas gdy w poprzedniej, 4. grupie znalazły się prawie same kobiety. Bardzo wysoki odsetek osób z grupy 5 określa swój poziom materialny jako dobry, co już stanowi pewną przesłankę wskazującą, że respondenci mogą pozytywnie zareagować na przekaz reklamowy. Analizując charakterystykę odpowiedzi

Tabela 1. Charakterystyka klastrów (skupień) uzyskanych za pomocą metody k-średnich z losowym wyborem środków skupień

Opis charakterystyki (zmiennej)	Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5	Populacja
liczba obserwacji						
	2255	1942	3931	1603	1631	11 362
płeć						
mężczyźni	89%	15%	29%	5%	98%	45%
kobiety	11%	85%	71%	95%	2%	55%
średni wiek respondenta						
	47 lat	45 lat	53 lata	31 lat	34 lata	45 lat
średni dochód netto						
	2203 zł	2026 zł	2494 zł	3358 zł	3250 zł	2587 zł
znajomość języka angielskiego						
zna język czynnie	2,8%	7,6%	1,7%	49,0%	38,9%	14,9%
zna język biernie	8,4%	12,8%	7,6%	24,1%	23,5%	13,3%
nie zna języka	88,8%	79,7%	90,7%	26,9%	37,6%	71,8%
korzystanie z Internetu						
tak	3%	11%	8%	82%	89%	30%
nie	97%	89%	92%	18%	11%	70%
ocena poziomu materialnego						
dobry	16,2%	10,2%	52,8%	67,7%	61,8%	41,7%
ani dobry, ani zły	38,0%	27,9%	35,3%	24,2%	26,5%	31,7%
zły	45,9%	61,9%	11,8%	8,0%	11,6%	26,6%
lubię mieć rzeczy, których inni mogą mi zazdrościć						
tak	42,0%	12,1%	3,6%	46,2%	25,7%	20,2%
nie	37,8%	71,8%	88,1%	34,1%	53,6%	64,5%
nie mam zdania	20,2%	16,1%	8,2%	19,7%	20,7%	15,4%
lubię kupować rzeczy, które nie mają praktycznego znaczenia						
tak	11,0%	4,8%	2,2%	66,6%	3,4%	7,0%
nie	80,2%	88,9%	95,4%	19,7%	90,3%	86,5%
nie mam zdania	8,8%	6,3%	2,4%	13,7%	6,4%	6,5%
samo robienie zakupów sprawia mi przyjemność						
tak	36,0%	55,5%	25,4%	78,7%	23,6%	39,9%
nie	42,7%	24,5%	56,2%	9,1%	56,3%	41,5%
nie mam zdania	21,3%	20,0%	18,5%	12,2%	20,1%	18,7%
zadowolenie z poziomu dóbr i usług						
zadowolony	80,4%	23,8%	96,9%	93%	78,5%	78%
niezadowolony	19,6%	76,2%	3,1%	7%	21,5%	22%

Źródło: opracowanie własne.

respondentów (mężczyźni, z miast, dobrze wykształceni) można wnioskować, że chociaż twierdzą co prawda, że nie lubią robić zakupów, to jednak biorąc pod uwagę ich wysokie dochody, warto skierować do tej grupy przekaz reklamowy. Wydaje się, że osoby z opisywanej grupy powinny zainteresować się propozycją rzeczy nowoczesnych, elektroniki, gadżetów, zwłaszcza biorąc jeszcze pod uwagę fakt, że są to osoby młode. Wydaje się, że skuteczne może być wysłanie przekazu reklamowego w wersji elektronicznej, ponieważ aż 89% osób z grupy korzysta z Internetu.

Stosując metodę k-średnich z losowym wyborem początkowych środków skupień otrzymaliśmy pięć klastrów o wyrazistej charakterystyce, gdy naszym celem było znalezienie grup osób, do których obiecujące wydaje się skierowanie odpowiednio opracowanych, różnych przekazów reklamowych. Z charakterystyki poszczególnych grup (pięciu klastrów) wywnioskowaliśmy, że do czterech uzasadnione wydaje się skierowanie przekazu reklamowego, a koszty tej działalności okażą się opłacalne. Jedna z uzyskanych grup, skupia osoby, które raczej nie będą zainteresowane przekazem i skierowanie przekazu wiązałoby się raczej tylko z poniesieniem niepotrzebnych kosztów. Zastosowanie metody k-średnich z losowym wyborem początkowych środków skupień dało klastry o zbliżonej liczbie obserwacji – klientów, w każdym. Nie było zmiennych, których wartości byłyby bardzo mocno zróżnicowane pomiędzy klastrami. Mogliśmy stwierdzić, że albo dana wartość zmiennej występowała w klastrze, lub nie występowała.

Zakończenie

Należy pamiętać, że metody analizy skupień powinny być traktowane tylko jako narzędzia pomocne w przeprowadzaniu segmentacji klientów na potrzeby prowadzenia kampanii reklamowych. Ostateczna decyzja będzie zawsze zależeć od kierujących prowadzeniem kampanii reklamowych i może spowodować konieczność uwzględnienia np. nowo pojawiających się zjawisk, czy też wydarzeń, które nie były brane pod uwagę przy tworzeniu segmentów klientów. Skierowanie przekazu reklamowego do pewnych osób może okazać się opłacalne, pomimo, że z otrzymanej segmentacji za pomocą analizy skupień może wynikać, że racjonalne byłoby przeciwne działanie: nie kierowanie przekazu. Rozważany problem prowadzenia kampanii reklamowych należy bowiem do problemów nieustrukturalizowanych, z elementami niepewności, niedokładności, subiektywizmu (można o tym wnioskować choćby na podstawie odpowiedzi respondentów na pytania ankiety).

Uzyskiwane wyniki dają możliwość pewnej elastyczności postępowania. Pomimo, że wyniki zastosowania metod analizy skupień nie mogą być traktowane jako dające bezwzględne, obligatoryjne rozstrzygnięcia decyzji w kampanii reklamowej, przeprowadzone przez nas rozważania teoretyczne i doświadczenia praktyczne wskazują, że ich stosowanie okazuje się być bardzo pomocne w podejmowaniu decyzji w prowadzeniu kampanii reklamowych. Przydatne mogą być różne metody. W decydującym stopniu rozstrzyga o tym cel zastosowania oraz charakter posiadanych danych.

W naszym przypadku, najbardziej przydatna i dająca najprecyzyjniejsze wskazówki okazała się metoda k-średnich z losowym wyborem początkowych środków skupień. Na-

tomiast jedyną metodą spośród przez nas analizowanych, która nie dała w naszym przypadku zadawalających wyników była metoda średniej grupowej. Klastry różniły się bardzo znacznie liczbą klientów, a szczegółowa charakterystyka poszczególnych segmentów nie dawała podstaw budowy ofert reklamowych ukierunkowanych dla grup klientów.

Interesujące może być prowadzenie dalszych badań, dotyczących zastosowania metod analizy skupień w prowadzeniu kampanii reklamowych. Zamierzamy podjąć próby zastosowania także innych metod, niż przedstawione w artykule, np. rozmytej analizy skupień oraz algorytmu maksymalizacji wartości oczekiwanej. Wartym uwagi może być podjęcie próby oszacowania zysków osiąganych dzięki oferowaniu różnych produktów i usług, klientom należącym do różnych segmentów, wydzielanych za pomocą różnych metod skupiania.

Bibliografia

Applied Analytics Using SAS Enterprise Miner 5.3. Course Notes, część I i część II, SAS Institute Inc. Cary, NC, USA, 2008.

Diagnoza Społeczna 2007: Warunki i jakość życia Polaków, www.diagnoza.com, dostęp 30 grudnia 2009.

Internetowy Podręcznik Statystyki, www.statsoft.pl/textbook, dostęp 30 grudnia 2009.

Jain A.K., Dubes R.C., *Algorithms for clustering data*, Prentice Hall, New Jersey 1988.

Jain A.K., Murty M.N., Flynn P.J., *Data Clustering: A Review*, ACM Computing Surveys, Vol. 31, No. 3, New York 1999.

Kohonen T., *Self-Organization and Associative Memory*, Springer-Verlag, Berlin 1984.

Kohonen T., *Self-Organizing Maps*, Springer, Berlin 2000.

Kotler P., *Marketing. Podręcznik europejski*, PWE, Warszawa 2002.

Kutera M., *Wykorzystanie metod analizy skupień w badaniach marketingowych*, praca magisterska napisana pod kierunkiem M. Lasek, Katedra Informatyki Gospodarczej i Analiz Ekonomicznych, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski, Warszawa 2010.

Lasek M., *Od danych do wiedzy. Metody i techniki* „Data Mining", Optimum, nr 2(22), 2004.

MacQueen J.B., *Some methods for classification and analysis of multivariate observations*, University of California Press, Berkeley 1967.

Michalski E., *Marketing. Podręcznik akademicki*, Wydawnictwo Naukowe PWN, Warszawa 2007.

Tryon R.C., *Cluster Analysis*, McGraw-Hill, New York 1970.

Clustering Methods Application for Customer Segmentation to Manage Advertisement Campaign

Summary

Clustering methods are recently so advanced elaborated algorithms for large collection data analysis that they have been already included today to data mining methods. Clustering methods are nowadays larger and larger group of methods, very quickly evolving and having more and more various applications. In the article, our research concerning usefulness of clustering methods in customer segmentation to manage advertisement campaign is presented. We introduce results obtained by using

four selected methods which have been chosen because their peculiarities suggested their applicability to our purposes. One of the analyzed method – k-means clustering with random selected initial cluster seeds gave very good results in customer segmentation to manage advertisement campaign and these results were presented in details in the article. In contrast one of the methods (hierarchical average linkage) was found useless in customer segmentation. Further investigations concerning benefits of clustering methods in customer segmentation to manage advertisement campaign is worth continuing, particularly that finding solutions in this field can give measurable profits for marketing activity.