JAK ZNALEŹĆ GRUPY PODOBNYCH KLIENTÓW, CZYLI METODY SEGMENTACJI

Grzegorz Migut
StatSoft Polska Sp. z o.o.

Wstęp

W dobie globalnej konkurencji i coraz szybciej zmieniających się preferencji klientów ogromne znaczenie ma nawiązywanie indywidualnych relacji pomiędzy dostawcą a osobami korzystającymi z jego usług. Tego typu relacje umożliwiają lepsze poznanie klienta, jego oczekiwań i preferencji, a co za tym idzie umożliwiają lepsze zaspokojenie jego potrzeb. Idealną sytuacją jest, gdy znamy wszystkich naszych klientów i potrafimy odpowiedzieć na ich indywidualne wymagania.

W takim przypadku posiadanie 100 000 klientów owocowałoby wyróżnieniem 100 000 profilów działania i zapewne przynosiłoby dobry efekt, jeśli chodzi o poziom zadowolenia klientów. Inną kwestią jest sens ekonomiczny tego typu strategii. Indywidualne traktowanie takiej liczby klientów pociągałoby za sobą kolosalne koszty i dezorganizowało pracę handlowców.

Indywidualne traktowanie olbrzymiej liczby klientów jest więc tak naprawdę mało realne. Realne jest natomiast takie działanie, by klienci mieli wrażenie, że są indywidualnie traktowani. Można to osiągnąć za pomocą odpowiednio przeprowadzonej segmentacji klientów.

Segmentacja polega na podziale niejednorodnej grupy obiektów (klientów) na grupy. Wszystkie osoby znajdujące się w tej samej grupie uważane są za podobne do siebie, osoby znajdujące się w różnych grupach są różne. Dzięki tego typu podziałowi nie musimy już określać tylu strategii, ilu mamy klientów. Wystarczy jeśli dany sposób postępowania przypiszemy do całej grupy (segmentu) podobnych osób. Liczba segmentów zależy od tego, jak zróżnicowani są nasi klienci, w praktyce jednak nie spotyka się firm, które z powodzeniem zarządzają i komunikują się z 10 lub więcej segmentami [3]. Oczywiście segmentacja ma sens jedynie wtedy, gdy planujemy wyszczególnione grupy traktować w odmienny sposób.

Innymi zaletami segmentacji jest możliwość patrzenia na naszych klientów z nieco innej perspektywy, niejako z lotu ptaka. Taka perspektywa może umożliwić lepsze zrozumienie ich cech i zachowań. Przyjęty schemat segmentacji może przyczynić się też do poprawy

komunikacji w zespole handlowców. Dzięki niemu mamy możliwość spójnie i precyzyjnie określić grupy klientów, którymi jesteśmy zainteresowani.

Wyniki przeprowadzonej segmentacji mogą również zasugerować działania, jakie należy podjąć w stosunku do określonych grup. Na przykład jeśli osoby znajdujące się w dwóch różnych segmentach mają bardzo podobne cechy demograficzne, a tylko jeden reprezentuje wartościowych klientów, istnieje szansa, że klienci z drugiego segmentu mogą stać się równie wartościowi, jeśli podejmiemy w stosunku do nich odpowiednie działania [5].

Wreszcie wyniki analiz są znakomitym punktem wyjścia do dalszych analiz data mining. Analizy takie (np. prognozy sprzedażowe) wykonujemy już nie dla całego niejednorodnego zbioru klientów, lecz co jest bardziej skuteczne i poprawne merytorycznie, dla wybranego jednorodnego segmentu.

Oczywiście dokonując segmentacji klienta, możemy brać pod uwagę wiele różnych związanych z nim aspektów. Różne rodzaje segmentacji zostały szczegółowo wymienione w artykule *Analiza danych i CRM - przegląd*, znajdującym się w niniejszym opracowaniu.

Metody data mining w segmentacji klientów

Podczas budowy modelu data mining zwykle podajemy na wstępie dane reprezentujące rozpoznane wcześniej wzorce (grupy), które model ma następnie odtwarzać i na tej podstawie klasyfikować do danej grupy nowe obserwacje. W przypadku segmentacji sytuacja jest inna. Na wstępie analizy nie dysponujemy żadną informacją, jakie segmenty występują w danych, ani też ile jest tych segmentów. Wiedzę tę pragniemy dopiero zdobyć w wyniku analizy. Tak sformułowane zadanie analityczne wymaga zastosowania jednej z metod nieukierunkowanego data mining. Najbardziej popularnymi metodami stosowanymi do segmentacji są metody analizy skupień oraz sieci neuronowe Kohonena (SOM).

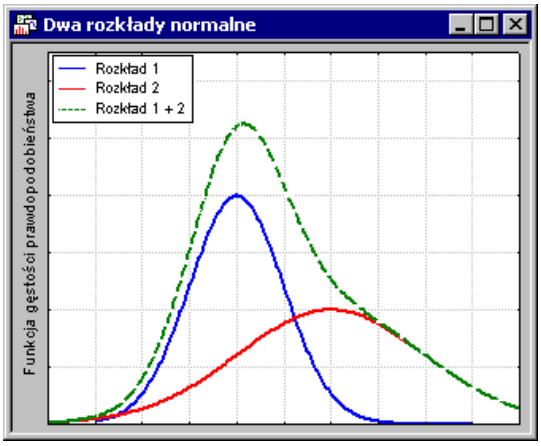
Analiza skupień - metoda k-średnich i EM

Celem **analizy skupień** (*cluster analysis*) jest wyodrębnienie ze zbioru danych obiektów, które byłyby podobne do siebie, i łączenie ich w grupy. W wyniku działania tej analizy z jednego niejednorodnego zbioru danych otrzymujemy grupę kilku jednorodnych zbiorów. Obiekty znajdujące się w tym samym zbiorze uznawane są za „podobne do siebie", obiekty z różnych zbiorów traktowane są jako „niepodobne". Pojęcie analizy skupień obejmuje faktycznie kilka różnych algorytmów klasyfikacji. Do najważniejszych należy zaliczyć metodę k-średnich oraz EM.

Stosowanie metody k-średnich wymaga od nas podania liczby grup, na które zostanie podzielony wejściowy zbiór danych. Jedna z wersji tej metody polega na losowym wyborze *k* obiektów z analizowanego zbioru i uznania ich za środki *k* grup. Każdy z pozostałych obiektów jest przypisywany do grupy o najbliższym mu środku. Następnie oblicza się nowe środki każdej podgrupy na podstawie średnich arytmetycznych ze współrzędnych zawartych w nich obiektów. W kolejnym kroku następuje przegrupowanie elementów grup, każdy obiekt jest przesuwany do tej grupy, do której środka ma najbliżej. Procedurę

tę powtarzamy do momentu, gdy w danej iteracji żaden z obiektów nie zmieni swojej podgrupy. Pewną wadą tej metody jest konieczność odgórnego określenia liczby skupień występujących w danych, dlatego też zaleca się powtórzenie procedury dla różnych wartości *k* i wybranie tej, dla której zbiór danych jest podzielony najlepiej.

Metoda EM jest czasem nazywana analizą skupień bazującą na prawdopodobieństwie lub statystyczną analizą skupień. Program wyznacza skupienia, zakładając różnorodne rozkłady prawdopodobieństwa zmiennych uwzględnianych w analizie. Na początku działania algorytmu, podobnie jak w metodzie k-średnich, musimy podać liczbę skupień, jakie powinny być wyodrębnione ze zbioru wejściowego.

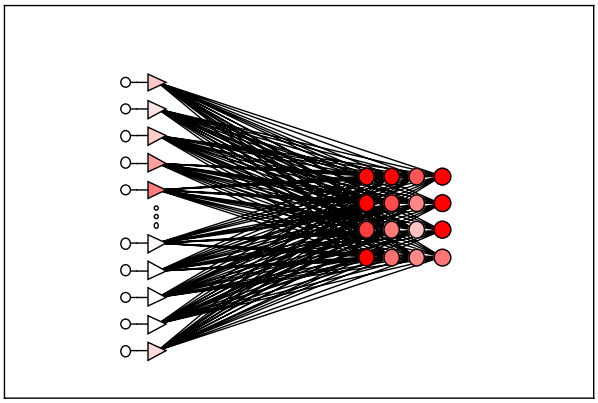


Załóżmy, że przeprowadziliśmy badania w pewnej dużej zbiorowości pod kątem jednej cechy ciągłej. Zaobserwowany rozkład tej cechy był zgodny z funkcją gęstości opisaną linią przerywaną (*Rozkład 1+2*) charakteryzującą się pewną średnią oraz odchyleniem standardowym. Wiemy też, że w zbiorowości tej występują dwa segmenty (na przykład kobiety i mężczyźni) o różnych parametrach funkcji gęstości w swoich segmentach. Algorytm EM ma na celu określenie parametrów rozkładów segmentów na podstawie rozkładu całej grupy oraz przydzielenie poszczególnych obserwacji do najbardziej odpowiadających im segmentów (klasyfikacja następuje na zasadzie prawdopodobieństwa). Na naszym rysunku rozkłady dwóch segmentów zostały przedstawione jako *Rozkład 1* oraz *Rozkład 2*. Po zsumowaniu dają one funkcję rozkładu całej zbiorowości (*Rozkład 1+2*). Algorytm EM dokonuje klasyfikacji nie tylko przy założeniu normalności rozkładu, jak to zaprezentowano na rysunku, wykorzystując go, można również określić inną funkcję gęstości dla badanej cechy (badanych cech).

Sieci neuronowe - sieć Kohonena (Self Organizing Map)

Sieć Kohonena (SOM) została zaprojektowana do uczenia w trybie bez nauczyciela – podczas uczenia ustalanie parametrów sieci nie jest sterowane za pomocą wartości wyjściowych, podczas nauki prezentowane są jedynie dane kierowane na wejścia sieci. Sieć ta posiada dwie warstwy: warstwę wejściową oraz warstwę wyjściową składającą się z neuronów radialnych. Warstwa ta znana jest również jako warstwa tworząca mapę topologiczną,

ponieważ takie jest jej najczęstsze zastosowanie. Neurony w warstwie tworzącej mapę topologiczną zwykle wyobrażamy sobie jako węzły dwuwymiarowej siatki, chociaż możliwe jest również tworzenie jednowymiarowych sieci w postaci długich łańcuchów.



Sieci Kohonena uczone są przy wykorzystaniu algorytmu iteracyjnego. Rozpoczynając od początkowych, wybranych w sposób losowy centrów radialnych, algorytm stopniowo modyfikuje je w taki sposób, aby odzwierciedlić skupienia występujące w danych uczących. Iteracyjna procedura ucząca dodatkowo porządkuje neurony reprezentujące centra położone blisko siebie na mapie topologicznej.

Podstawowy iteracyjny algorytm działa przez dużą liczbę epok (podczas każdej epoki prezentowany jest sieci cały zestaw danych) w następujący sposób [6]:

- pokazywany jest zestaw danych wejściowych ze zbioru uczącego,
- wszystkie neurony sieci wyznaczają swoje sygnały wyjściowe, stanowiące odpowiedź na podane wejścia,
- wybierany jest neuron zwycięski (tzn. ten, który reprezentuje centrum najbardziej zbliżone do prezentowanego na wejściu przypadku),
- neuron zwycięski modyfikowany jest w taki sposób, aby upodobnić jego wzorzec do prezentowanego przypadku. W tym celu wyznaczana jest ważona suma przechowywanego w neuronie centrum oraz przypadku uczącego,
- wraz ze zwycięskim neuronem w podobny sposób modyfikowane są parametry jego sąsiadów (sąsiedzi wyznaczani są w oparciu o przyjęty wzór topologii sieci).

Algorytm wykorzystuje zmienny w czasie współczynnik uczenia, który służy do wyznaczania ważonej sumy i powoduje, że zmiany początkowo duże i szybkie - stają się coraz bardziej subtelne w trakcie kolejnych epok. Umożliwia to ustalenie centrów w taki sposób, że stanowią one pewien kompromis pomiędzy wieloma przypadkami powodującymi zwycięstwo rozważanego neuronu.

Własność uporządkowania topologicznego jest osiągana przez zastosowanie w algorytmie koncepcji sąsiedztwa. Sąsiedztwo stanowią neurony otaczające neuron zwycięski. Sąsiedztwo, podobnie jak współczynnik uczenia, zmniejszane jest wraz z upływem czasu, tak więc

początkowo do sąsiedztwa należy stosunkowo duża liczba neuronów; w końcowych etapach sąsiedztwo ma zerowy zasięg. Ma to istotne znaczenie, ponieważ w algorytmie Kohonena modyfikacja wag jest w rzeczywistości przeprowadzana nie tylko w odniesieniu do neuronu zwycięskiego, ale również we wszystkich neuronach należących do sąsiedztwa.

Po nauczeniu sieci Kohonena poprawnego rozpoznawania struktury prezentowanych danych można jej użyć jako narzędzia przeprowadzającego wizualizację danych w celu ich lepszego poznania. Ważnym elementem przygotowania sieci Kohonena do bieżącego użytkowania jest właściwe opisanie uformowanej mapy topologicznej. Ustalenie związków pomiędzy skupieniami a znaczeniami wymaga zwykle odwołania się do dziedziny, której dotyczy analiza [6].

Przykład segmentacji klientów

Jako przykład wykorzystania analiz typu *data mining* do analizy danych zaprezentowany zostanie przykład segmentacji behawioralnej (opartej na danych opisujących zachowanie klienta - w tym przypadku były to dane transakcyjne), którą przeprowadzono za pomocą technik analizy skupień. Dodatkowo w celu walidacji wyników użyto sieci Kohonena.

Podstawą przeprowadzonej analizy były informacje pochodzące z fragmentu bazy danych udostępnionej przez jeden z banków komercyjnych z Europy Środkowej. Udostępniona baza zawierała szereg informacji zarówno demograficznych, jak i transakcyjnych. Zawierała między innymi zapis wszystkich operacji (wpłaty, wypłaty, przelewy) dokonywanych przez klientów banku w pewnym okresie czasu, a także informacje dotyczące zaciąganych kredytów (historia spłat itp.).

Do analizy wykorzystano jedynie informacje dotyczące transakcji. Zawarte były one w jednej tabeli o nazwie *Transaction*, w której przechowywano ponad 1 000 000 rekordów. Każdy z rekordów był zapisem transakcji przeprowadzonej przez jednego z 4 500 klientów banku uwzględnionego w bazie.

Celem analizy było wykrycie ewentualnych segmentów wśród klientów banku oraz określenie cech osób należących do poszczególnych grup. Na tej podstawie można rozwijać określone działania marketingowe skierowane do poszczególnych grup, dostosować ofertę do modelu zachowania klienta. Wyniki analizy mogły także powiększyć wiedzę na temat naszych klientów. Ponieważ zgromadzone dane nie zawierały żadnego podziału znanego z góry, innymi słowy w danych nie można było wyróżnić zmiennej zależnej, dlatego też do analizy zostały wykorzystane metody należące do grupy metod nieukierunkowanego *data mining*: analiza skupień i sieci neuronowe (SOM).

Przygotowanie danych do analizy

Pierwszym etapem analizy był tak zwany *preprocessing,* czyli wstępna analiza danych mająca na celu takie przekształcenie pierwotnego zbioru, by można go było jak najprościej przeanalizować w zasadniczej części analizy. W poniższym przykładzie dane zostały przygotowane w następujący sposób.

Na podstawie trzech kolumn tabeli bazy danych *transaction - k_symbol, type, operation* określono rodzaje transakcji dokonywane przez klientów w ten sposób, że każdą kombinację wartości tych kolumn potraktowano jako jeden rodzaj transakcji. W danych zaobserwowano 14 różnych kombinacji - 14 różnego rodzaju transakcji.

- *WplataOdsetki* - Wpłata przez bank odsetek od kwoty na rachunku,
- *WplataPrzelew* - Wpłata na konto przelewem,
- *WplataGotowka* - Wpłata gotówki,
- *WyplataPrzelewGosp* - Wypłata przelewem w celu uiszczenia opłat związanych z utrzymaniem gospodarstwa domowego,
- *WyplataGotowka* - Wypłata gotówki,
- *WyplataGotowkaOpBank* - Opłata za wyciągi bankowe itp.,
- *WyplataPrzelewPoz* - Wypłata przelewem w celu spłacenia raty kredytu lub pożyczki,
- *WyplataPrzelewInne* - Wypłata przelewem na inne cele,
- *WyplataPrzelewUbezp* - Wypłata przelewem opłata ubezpieczeniowa,
- *WplataPrzelewEmer* - Wpłata przelewem emerytury lub renty,
- *WyplataKartaKred* - Wypłata przy pomocy karty kredytowej,
- *WyplataGotowkaGosp* - Wypłata gotówki w celu uiszczenia opłat związanych z utrzymaniem gospodarstwa domowego,
- *WyplataKarneOds* - Potrącenie przez bank odsetek za debet,
- *WyplataGotowkaUbezp* - Wypłata gotówki na opłatę ubezpieczeniową.

Dodatkowo z bazy pobrano informację odnośnie numeru konta przypisanego do transakcji oraz kwoty, na jaką została dokonana. Arkusz zawierał więc trzy zmienne - *typ transakcji*, *numer konta* i *kwota* oraz ponad 1 000 000 przypadków. Pierwsze zadanie polegało na takim zwinięciu arkusza danych, by każdemu klientowi odpowiadał dokładnie jeden wiersz arkusza (przypadek). Zwinięcia dokonano w ten sposób, że każdy rodzaj transakcji został zamieniony w zmienną, która przyjmowała wartości równe sumie wszystkich transakcji tego samego typu dokonanych przez danego klienta. Przekształcenie to wykonano za pomocą opcji Zwiń/Rozwiń (*Unstacking/Stacking)*. W ten sposób otrzymano 14 zmiennych zawierających kwoty poszczególnych transakcji.

Podobną operację przeprowadzono dla ilości transakcji (po przekształceniu wartość zawarta w każdej komórce była sumą ilości transakcji jednego rodzaju wykonanych przez danego klienta) przez co otrzymano kolejne 14 zmiennych.

Przed przystąpieniem do zasadniczej części analizy sprawdzono jeszcze rozkłady poszczególnych zmiennych. Analizując uzyskane wyniki, zauważono, że zmienna *WyplataGotowkaUbezSum* oraz odpowiadająca jej zmienna *WyplataGotowkaUbezpTrans* przyjmuje wartość zero dla wszystkich przypadków za wyjątkiem jednego. Taki rozkład powodował, że zmienne te nie zawierały w sobie żadnej informacji, a uwzględnienie ich w analizie mogło jedynie zaburzyć uzyskane wyniki (zmienna ta mogła być efektem błędnych zapisów w bazie transakcyjnej) W dalszej analizie nie uwzględniano tej pary zmiennych.

Kolejna para zmiennych, które poddano ocenie to *WyplataGotowkaGospSum* oraz *WyplataGotowkaGospTran* określające ilość i kwotę transakcji wypłat gotówkowych na utrzymanie gospodarstwa domowego (rachunki itp.) W wypadku tych zmiennych, jedynie 78 przypadków przyjmowało wartość niezerową, tak więc ich wartość informacyjna była równie znikoma. Warto się również zastanowić nad sensem merytorycznym tych zmiennych. W praktyce, tak naprawdę, trudno sobie wyobrazić sytuację, w której wypłacamy gotówkę w kasie lub bankomacie i podajemy cel tej wypłaty. Z powyższych powodów również te dwie zmienne nie zostały uwzględnione w dalszej analizie.

Przekształcony arkusz danych zawierał 4 500 przypadków. Każdy przypadek składał się z 25 zmiennych:

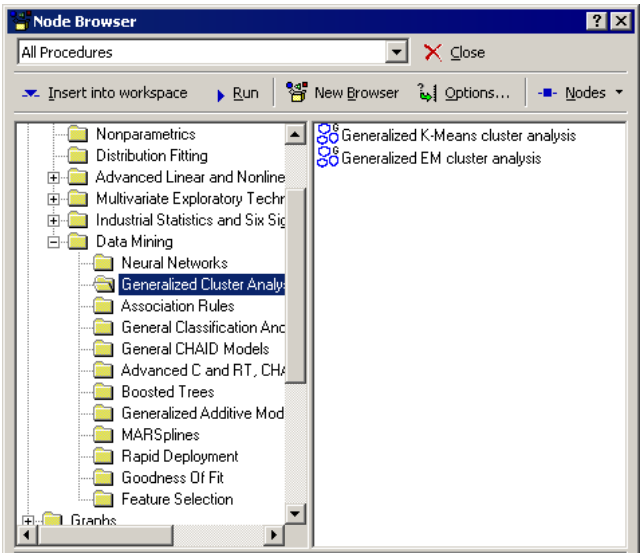
- 12 zmiennych określających łączną kwotę, na jaką zrealizowano danego rodzaju transakcję,
- 12 zmiennych opisujących ilość transakcji danego rodzaju,
- 1 zmienna identyfikująca klienta.

Wszystkie zmienne opisujące transakcje były zmiennymi liczbowymi.

Proces analizy

Wszystkie analizy przeprowadzone zostały w środowisku *STATISTICA Data Miner*. Najwygodniej było przeprowadzić je w specjalnie zaprojektowanej przestrzeni roboczej. Umieszczono w niej arkusz wejściowy *Transakcje.sta* i wybrano zmienne biorące udział w analizie. Wszystkie zmienne opisujące transakcje oznaczono jako predyktory ciągłe, pominięto natomiast zmienną identyfikującą klienta.

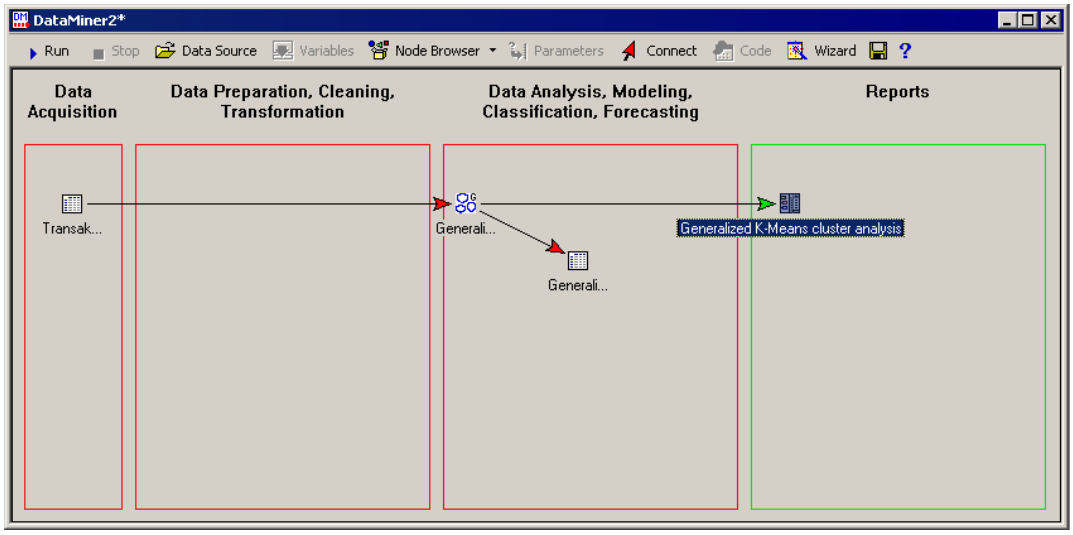
Segmentacji dokonano przy pomocy uogólnionej metody k-średnich *(Generalized K-Means cluster analysis)*. Z przeglądarki węzłów wybrano odpowiedni węzeł i umieszczono go w przestrzeni roboczej.



Dużym mankamentem wybranej metody jest fakt, iż musimy na wstępie analizy określić liczbę segmentów, na które algorytm ma podzielić zbiór wejściowy. W praktyce jednak zwykle nie znamy tej liczby, pragniemy dopiero odnaleźć najlepsze rozwiązanie. Niedogodność tę można ominąć, stosując zaimplementowany w metodzie sprawdzian krzyżowy. Dzięki sprawdzianowi krzyżowemu można wyznaczyć i ocenić najlepszy układ skupień - program automatycznie określa najbardziej odpowiednią liczbę skupień (segmentów). Dlatego też zaznaczono opcję V-*Fold Crossvalidation*.

Warto zauważyć, że wybrana metoda umożliwia zapisanie zbudowanego modelu w postaci kodu C i PMML oraz *STATISTICA Visual Basic*, dzięki czemu możemy z łatwością stosować zbudowany model dla nowych danych zarówno w środowisku *STATISTICA*, jak i poza nim.

Wybrany moduł podłączono do wejściowego arkusza danych, a następnie uruchomiono proces analizy, wybierając polecenie *Run*. Po wykonaniu analizy jej wyniki zostały umieszczone w węzłach wynikowych utworzonych dla tej metody.



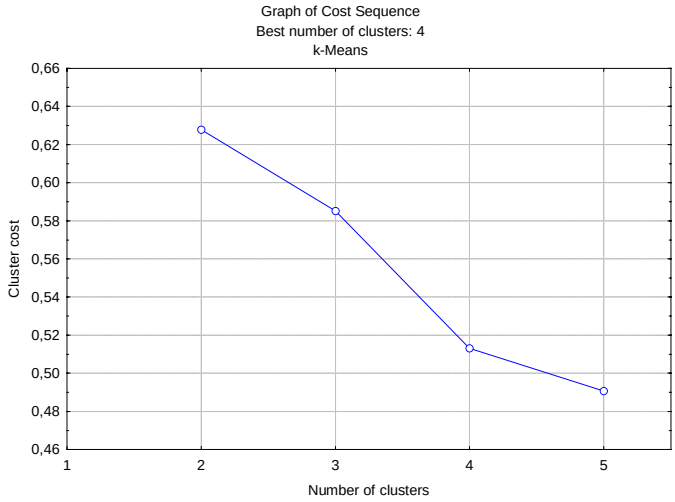
Analiza uzyskanych wyników

Dla wybranej metody wygenerowane zostały dwa węzły wynikowe (widoczne na rysunku powyżej). Pierwszy z nich zawierał pierwotny arkusz danych z nową zmienną określającą, do którego segmentu przydzieleni zostali poszczególni klienci. Drugi węzeł zawierał raporty przedstawiające szczegółowe wyniki analizy. Analizując te wyniki można zauważyć, że algorytm zidentyfikował cztery segmenty Poniżej zamieszczono raport wyników.

	Summary for k-means clustering (Transakcje) Number of clusters: 4 Total number of training cases: 4500
Algorithm	k-Means
Distance method	Euclidean distances
Initial centers	Maximize initial distance
MD casewise deletion	Yes
Cross-validation	3 folds
Testing sample	0
Training cases	4500
Training error	0.509791
Number of clusters	4

W węźle raportów znajdował się także arkusz wskazujący odległość środków poszczególnych segmentów od pozostałych środków segmentów, a także arkusz zawierający środki segmentów dla poszczególnych zmiennych, który wykorzystano w dalszej analizie.

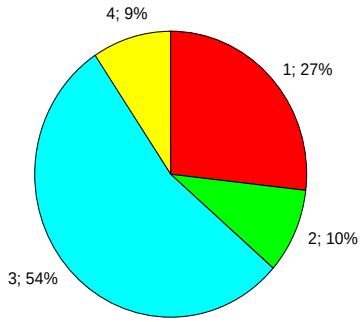
Ostatnim elementem węzła raportowego był tak zwany wykres osypiska przedstawiający wynik działania mechanizmu walidacji krzyżowej. Wykorzystując test krzyżowy, algorytm dzielił zbiór wejściowy kolejno na coraz większą liczbę segmentów, a następnie sprawdzał, jaka jest precyzja podziału dla każdego z nich. Dla metody *k-średnich* miarą precyzji podziału jest przeciętna odległość elementów zbioru wejściowego od środka segmentu, w jakim się znajdują.



Analizując wykres, można zauważyć znaczną poprawę precyzji podziału przy zwiększeniu liczby segmentów z trzech do czterech. Dodając jeszcze jeden segment uzyskuje się już znacznie mniejszą poprawę precyzji, stąd za optymalną liczbę segmentów należy uznać cztery.

W obrębie wykorzystywanej przestrzeni roboczej przeprowadzono dodatkowe analizy, wybierając interesującą metodę z przeglądarki węzłów i łącząc ją z analizowanym zbiorem danych.

Na podstawie arkusza wyników uzyskanego w wyniku analizy sprawdzono najpierw liczność poszczególnych segmentów. Z przeglądarki węzłów wybrano wykres kołowy (*Pie charts*) i połączono go z arkuszem wynikowym. Wykres wygenerowano dla zmiennej *final classification* informującej, do którego segmentu trafiła dana obserwacja.



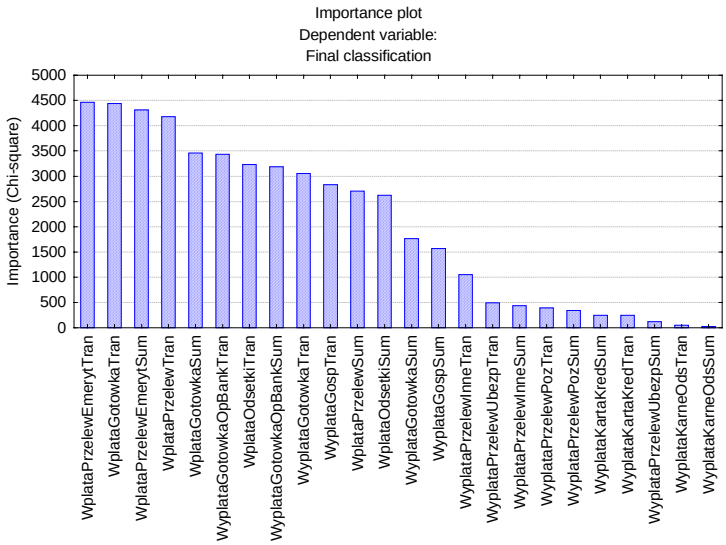
Na wykresie widać, że liczba klientów w poszczególnych segmentach przedstawia się następująco:

- pierwszy segment - 27 %,
- drugi segment - 10 %,
- trzeci segment - 54%,
- czwarty segment - 9%.

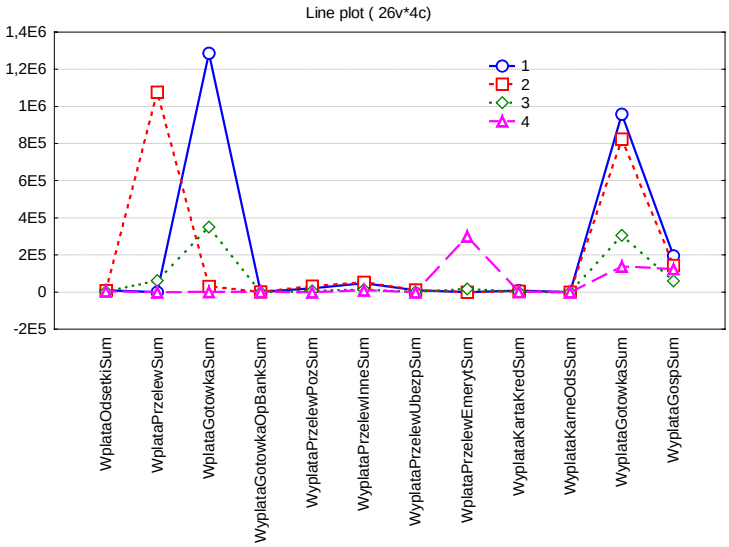
Na podstawie tego samego arkusza wyników zbadano, jakie zmienne miały największy wpływ na proces podziału. W tym celu z przeglądarki węzłów wybrano węzeł *Feature Selection and Root Cause Analysis.* W węźle tym zaznaczono dodatkowo opcję nakazującą wygenerowanie wszystkich możliwych raportów. Wybrany węzeł umieszczono w przestrzeni roboczej i połączono z arkuszem danych powstałym w wyniku analizy. Następnie dokonano specyfikacji zmiennych do analizy. Zmienna określająca segment, do którego należą poszczególni klienci, została określona jako skategoryzowana zmienna zależna. Wszystkie pozostałe zmienne (poza ID klienta) określono jako predyktory ciągłe.

Wyniki węzła *Feature Selection* zostały przedstawione na poniższym histogramie opisującym wpływ poszczególnych zmiennych na proces segmentacji. Można zauważyć, że największy wpływ na proces podziału miały: zmienna określająca ilość transakcji wpłaty gotówki, zmienne określające ilość transakcji i sumę gotówki wpłat przelewem emerytury lub renty oraz ilość transakcji dotyczących wpłat dokonanych przelewem (rys. poniżej).

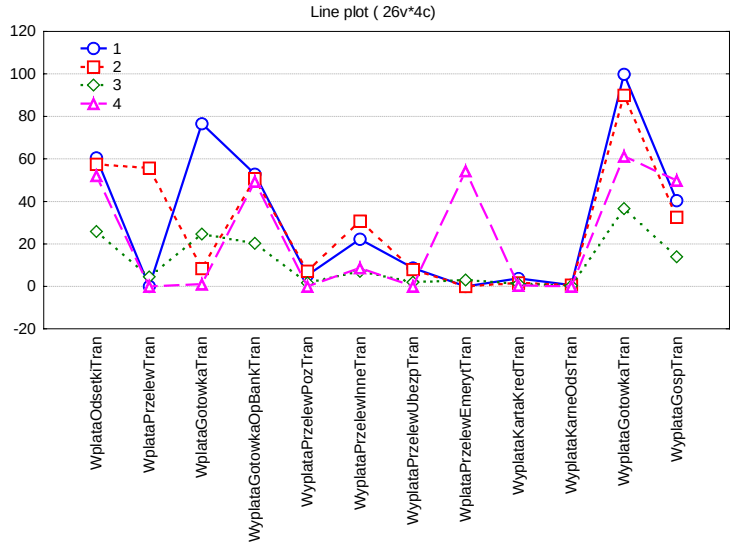
Bardzo pomocne w określaniu specyfiki poszczególnych segmentów są procedury graficzne. Skorzystano z nich, analizując jeden z arkuszy wyników analizy skupień zawierający środki segmentów dla poszczególnych zmiennych. Dzięki temu możliwe było określenie najbardziej typowych operacji dla poszczególnych segmentów.



Zawarty w węźle raportów arkusz oznaczono jako aktywny arkusz wejściowy (*Active dataset*), a następnie umieszczono w przestrzeni roboczej. Do analizy uzyskanego arkusza zastosowano wykres liniowy dla przypadków *Line Plots (case profiles)*. Warto zauważyć, że zakres wartości zmiennych opisujących ilość transakcji jest o wiele mniejszy od zakresu wartości zmiennych opisujących kwoty, na jakie dokonano transakcji. Różnica zakresów może spowodować, że zmienność w ilościach dokonywanych transakcji nie będzie zauważalna (zmienne te przyjmują o wiele mniejsze wartości). Rozwiązaniem tego problemu może być standaryzacja ujednolicająca zakresy wartości wszystkich zmiennych, bądź przeanalizowanie obu grup zmiennych na osobnych wykresach. W celu utrzymania przejrzystości wykresów wybrano to drugie rozwiązanie.



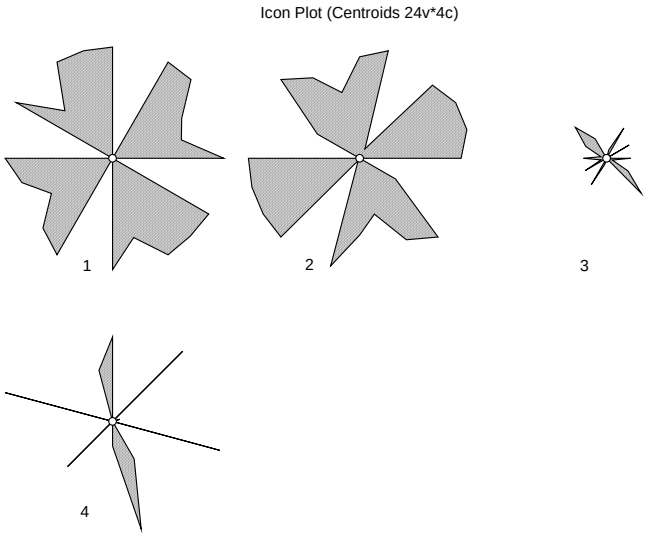
Powyższy wykres przedstawia środki poszczególnych skupień, który wygenerowano dla zmiennych opisujących kwoty transakcji. Można zauważyć, że pierwsza grupa to osoby dokonujące wpłat i wypłat w gotówce na największą kwotę. Druga grupa to osoby korzystające przede wszystkim z przelewów, jeśli chodzi o wpłacanie środków, natomiast wypłacające przede wszystkim w gotówce, w ilościach zbliżonych do segmentu pierwszego. Trzecia grupa to osoby głównie wpłacające i wypłacające gotówkę, jednak znacząco niższe kwoty od dwóch pierwszych grup. Czwarta grupa jest podobna do trzeciej jeśli chodzi o kwotę pieniędzy, którymi obraca, różni się od wszystkich źródłem wpłat, są to przelewy z tytułu renty lub emerytury.



Analizując zmienne opisujące ilość transakcji, można zauważyć, że najmniej aktywnymi klientami są osoby znajdujące się w trzecim segmencie. Najwięcej transakcji dokonują klienci z segmentów pierwszego i drugiego, nieco mniej z segmentu czwartego - czyli emeryci.

Niestety najbardziej liczni są klienci najmniej atrakcyjni; stanowią ponad połowę wszystkich klientów banku. Najbardziej atrakcyjni klienci z pierwszej i drugiej grupy stanowią w sumie 36% klientów.

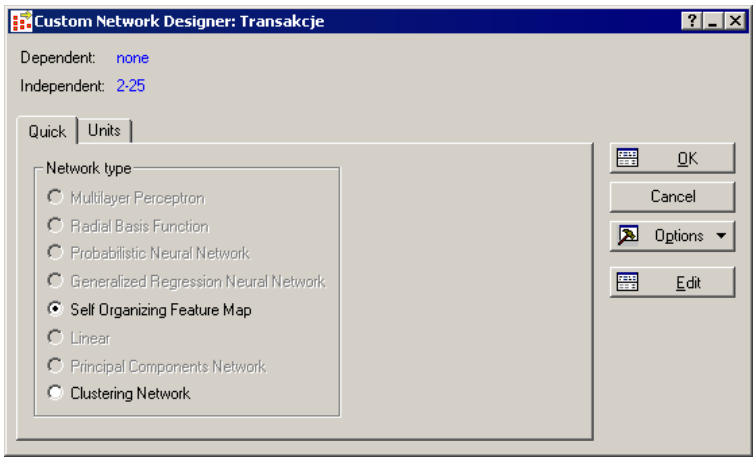
Kolejną metodą graficzną, jaką można wykorzystać do scharakteryzowania poszczególnych segmentów, jest wykres obrazkowy (*icon plots*). Na wykresie tego typu każdy segment prezentowany jest w formie jednego obrazka (na przykład twarzy Chernoffa, linii czy wykresów kołowych). Poniżej przedstawiono wykres obrazkowy, na którym poszczególne segmenty zaprezentowane zostały w formie wielokątów. Analizując go, można zauważyć znaczne podobieństwo pomiędzy osobami z segmentu pierwszego i drugiego. Osoby z trzeciego segmentu znacznie różnią się od pozostałych.



Sieci neuronowe Kohonena w segmentacji klientów

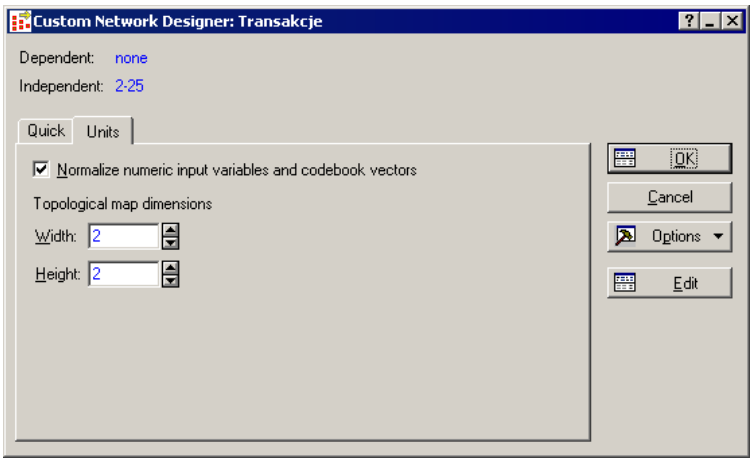
Sieci neuronowe SOM są konkurencyjną w stosunku do technik analizy skupień metodą segmentacji. W niniejszym przykładzie wykorzystano je do weryfikacji poprawności podziału przeprowadzonego za pomocą metody k-średnich.

Analizę przeprowadzono dla tego samego arkusza danych, tym razem w interakcyjnym środowisku *STATISTICA*. W menu Statystyki (*Statistics*) wybrano opcję Sieci Neuronowe (*Neural Networks*), w wyświetlonym oknie określono problem analityczny jako analizę skupień (*Cluster Analysis*). Wszystkie zmienne określono jako wejścia. Ponieważ sieci Kohonena pracują w trybie bez nauczyciela (nieukierunkowany data mining) nie określono żadnej zmiennej wyjściowej.

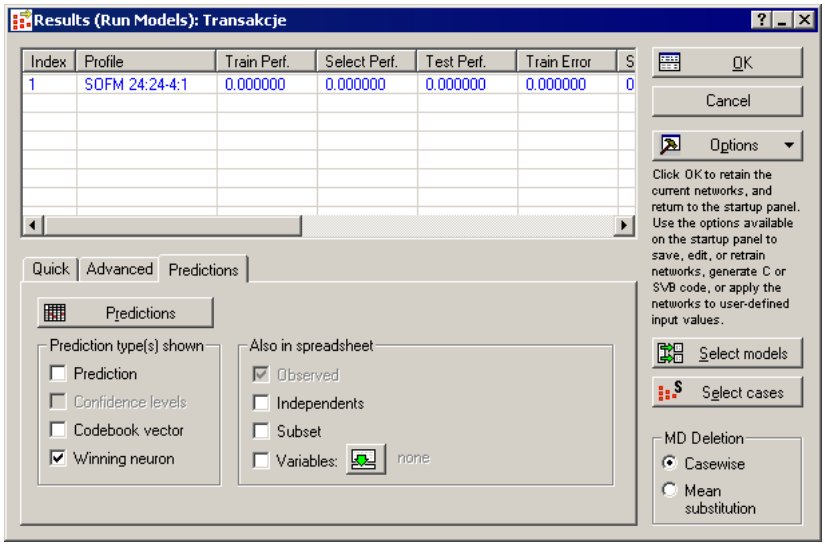


Po zatwierdzeniu ustawień, w nowym oknie dialogowym określono rodzaj oraz wielkość sieci. Na zakładce *Quick* (rys. powyżej) zaznaczono *Self Organizing Feature Map* (dla określonego wcześniej układu zmiennych dostępnymi topologiami były sieć grupującą *Clustering Network* oraz SOM).

Na zakładce *Units* określamy ilość neuronów wyjściowych. W sieci SOM każdy neuron wyjściowy reprezentuje jedno skupienie. Ponieważ algorytm testu krzyżowego wskazał na cztery skupienia, również w tym przypadku wybrano tę liczbę. Budowana mapa topologiczna miała więc wymiary 2x2. Pozostałe parametry analizy pozostawiono w domyślnym położeniu.

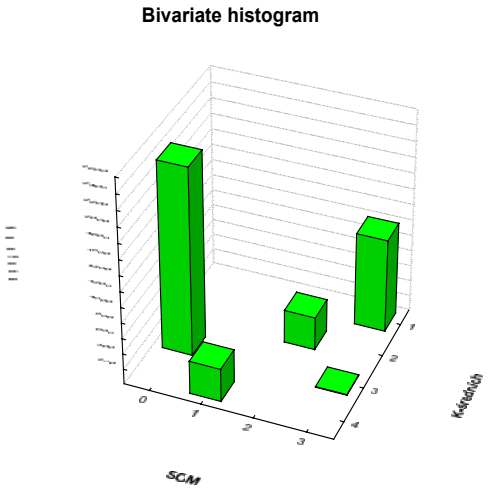


Po wykonaniu analizy kolejnym zadaniem było porównanie wyników sieci z podziałem określonym za pomocą analizy skupień. W tym celu przy pomocy opcji *Predictions* wygenerowano arkusz zawierający odpowiedzi sieci dla wszystkich przypadków analizowanego zbioru.



Powstały w ten sposób arkusz połączono za pomocą opcji *Merge* z arkuszem zawierającym wynik przeprowadzonej wcześniej analizy skupień. Odpowiedzi uzyskane za pomocą obydwu metod porównano za pomocą trójwymiarowego histogramu (*Bivariate Histograms*).

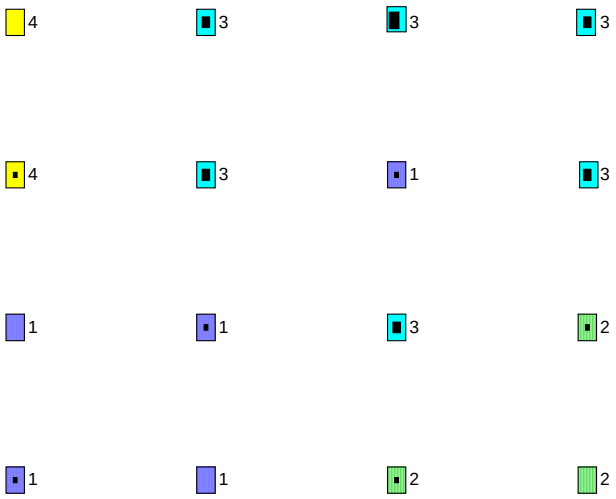
Analizując poniższy wykres, można zauważyć niemal pełną zgodność uzyskanych wyników (różnią się jedynie numery, jakie otrzymały poszczególne skupienia). Bardzo niewielka liczba obserwacji została odmiennie zaklasyfikowana przez obie metody.



Dużą zaletą sieci Kohonena jest możliwość redukcji wymiaru danych wejściowych do dwóch wymiarów, dzięki czemu możemy uzyskać prostą w interpretacji mapę topologiczną i na tej podstawie określić, które skupienia są podobne do siebie (dane dwa skupienia są tym bardziej do siebie podobne, im bliżej znajdują się na mapie topologicznej). Ustalona podczas analizy mapa o wymiarach 2x2 jest jednak zbyt mała, by móc wyciągać na jej podstawie jakiekolwiek wnioski dotyczące podobieństwa pomiędzy segmentami (wszystkie neurony są swoimi sąsiadami). Dlatego ostatnia z przeprowadzonych analiz była próbą budowy nieco większej sieci składającej się z 16 neuronów rozmieszczonych na mapie o wymiarach 4x4.

Analizę wykonano dla identycznych ustawień jak dla sieci 2x2. W wyniku analizy otrzymano mapę. Ostatnim krokiem było zaetykietowanie poszczególnych neuronów (którą grupę reprezentuje dany neuron) oraz dostosowanie jej wyglądu (określenie kolorów poszczególnych komórek).

Podobieństwo pomiędzy segmentem pierwszym i drugim stwierdzone podczas analizy metodą k-średnich jest również widoczne na mapie topologicznej – neurony reprezentujące obie grupy znajdują się w swoim bliskim sąsiedztwie. Najmniejsze podobieństwo wstępuje pomiędzy segmentem czwartym oraz drugim – leżą po przeciwnych częściach mapy.



Zbudowany model sieci możemy bardzo łatwo stosować dla nowych danych. I to zarówno w środowisku *STATISTICA* (po prostu uruchamiamy zbudowany model dla nowych danych), jak i poza nim, wykorzystując możliwość generowania kodu źródłowego. Warta uwagi jest też możliwość zapisania w pliku zarówno całego projektu *data mining*, jak i zbudowanego modelu sieci i udostępnienia ich innym użytkownikom *STATISTICA Data Miner*.

Literatura:

1. Berka P, *Guide to the financial Data Set Laboratory for Intelligent Systems*, University of Economics, Czech Republic.
2. Berry M., Gordon L., *Mastering Data Mining. The Art and Science of Customer Relationship Management*, John Wiley & Sons, Inc, New York 2000.
3. Berson A., Smith S., Thearling K., *Building Data Mining Applications for CRM*, McGraw Hill, New York 2000.
4. Hotho A, Maedche A. *Efficient Discovery of Client Profiles from a Financial Database*, Institute AIFB, Karlsruhe University, Germany.
5. Kumar S., *Customer Segmentation: Think Beyond What You Can See*, DM Direct Special Report, 2004.
6. *STATISTICA Neural Networks PL. Wprowadzenie do sieci neuronowych*, StatSoft Polska, 2001.