



ANALIZA SKUPIEŃ NA PRZYKŁADZIE SEGMENTACJI NOWOTWORÓW

Grzegorz Harańczyk, StatSoft Polska Sp. z o.o.

Analiza skupień to jedna z najbardziej znanych metod data miningu. Zaprezentujemy zastosowanie tej metody do segmentacji nowotworów, wykorzystując algorytm k-średnich i jego implementację w programie *STATISTICA*.

Wprowadzenie do analizy skupień

Ogólny problem badaczy wielu dyscyplin polega na organizowaniu obserwowanych danych w sensowne struktury lub grupowaniu danych. Obecnie zagadnienie to jest szczególnie istotne, gdyż coraz częściej mamy do czynienia z ogromnymi ilościami danych. Właśnie do tych celów można zastosować analizę skupień.

Analiza skupień (ang. *cluster analysis*, termin wprowadzony w 1939 roku przez Tryona), nazywana również segmentacją lub klastrowaniem danych, jest przykładem analizy polegającej na szukaniu i wyodrębnieniu z danych skupień, czyli grup obiektów podobnych. Jest to metoda nieukierunkowana (*unsupervised*), to znaczy, że wszelkie związki i prawidłowości znajdowane są tylko na postawie cech wejściowych.

Celem segmentacji jest wydzielenie grup obserwacji podobnych, dalszym etapem może być szukanie cech charakterystycznych dla obserwacji wchodzących w skład danej grupy. W przeciwieństwie do klasyfikacji wzorcowej (analizy z nauczycielem), polegającej na przyporządkowywaniu przypadków do jednej z określonych klas, tu klasy nie są znane ani w żaden sposób scharakteryzowane przed przystąpieniem do analizy. Jednak po scharakteryzowaniu wyodrębnionych skupień można w dalszym etapie badań klasyfikować nowe przypadki, przyporządkowując je do odpowiedniego skupienia.

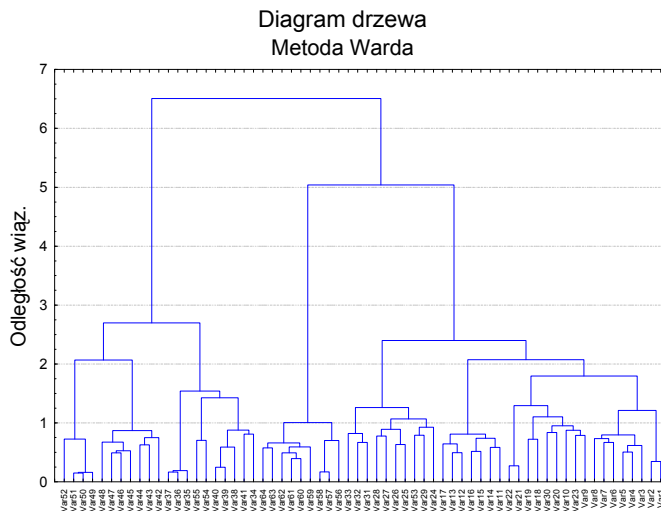
Pośrednio celem analizy skupień jest także weryfikacja jednorodności danych. Jeśli można wyróżnić skupienia, wtedy oczywiście danych nie można uznać za jednorodne.

Organizowanie obiektów w skupienia opiera się na szukaniu obserwacji podobnych. Aby móc porównywać obserwacje między sobą, określać, na ile są one do siebie podobne, musimy wprowadzić miarę podobieństwa obserwacji. W przypadku zmiennych jakościowych będą to tak zwane indeksy podobieństwa (np. indeks Russela i Rao, indeks Jaccarda,

indeks Sokala i Michnera), a w przypadku zmiennych ilościowych – odległości (np. odległość euklidesowa, odległość Czebyszewa, odległość Manhattan). Są też specjalne miary podobieństwa, które można stosować, gdy podczas analizy wykorzystujemy jednocześnie cechy o charakterze jakościowym i ilościowym.

Wyróżnia się dwa zasadnicze typy algorytmów grupowania danych: algorytmy hierarchiczne i algorytmy niehierarchiczne. Hierarchiczne metody aglomeracyjne prowadzą do stworzenia tzw. hierarchii drzewkowej elementów analizowanego zbioru (dendrogramu). Na wstępie procedury przyjmuje się, że każdy obiekt stanowi osobne skupienie, następnie krokowo łączy się w podzbiory podgrupy najbardziej do siebie podobne, aż do otrzymania jednego skupienia zawierającego wszystkie obserwacje. W ten sposób otrzymuje się wynikową segmentację, będącą uporządkowanym zestawieniem podziałów na segmenty.

Hierarchiczne metody grupowania nie wymagają wcześniejszego podania liczby skupień (na dendrogramach wyboru liczby skupień można dokonać na końcu analizy, przecinając go na odpowiedniej wysokości, rys. 1), ale wymagają dużej mocy obliczeniowej. Dla zbiorów danych o znacznej wielkości obliczenia mogą zająć dużo czasu lub wręcz być niewykonalne.



Rys. 1. Dendrogram

Metody niehierarchiczne są szybkie, ale wymagają wcześniejszego podania liczby skupień, do których dane mają zostać zakwalifikowane. Wybór liczby skupień ma duży wpływ na jakość uzyskanej segmentacji. Podanie zbyt dużej liczby skupień może spowodować, że wyznaczone skupienia będą co prawda wewnątrznie jednorodne, jednak utrudniona będzie interpretacja uzyskanych wyników i stosowanie ich w praktyce. Z drugiej strony, im mniejsza liczba skupień, tym skupienia są mniej jednorodne wewnątrznie. Za wadę może być również uznane to, że wewnątrz skupień nie mamy żadnego porządku, a także fakt, że gdy zmienimy liczbę skupień, na przykład zwiększymy o 1, to skupienia utworzone w wyniku nowego podziału nie będą zawierać się we wcześniej uzyskanych.



W niniejszym artykule przede wszystkim skupimy się na jednym z algorytmów z grupy metod niehierarchicznych – procedurze k-średnich. Jest to jeden z najpopularniejszych algorytmów analizy skupień.

W następnej części opiszemy algorytm k-średnich, a potem zastosujemy jego implementację w systemie *STATISTICA* do segmentacji nowotworów na podstawie poziomej ekspresji genów. Zaprezentujemy rozwiązanie problemu segmentacji opartego na przykładzie z książki Hastie, Tibshirani, Friedman [1].

Algorytm k-średnich

Standardowo algorytm k-średnich wymaga, aby wszystkie zmienne, użyte podczas analizy, były zmiennymi ilościowymi, a więc podobieństwo między obserwacjami będzie mierzone za pomocą odległości. Implementacja algorytmu k-średnich w *STATISTICA* (*Analiza skupień uogólnioną metodą k-średnich*) pozwala na wykorzystanie podczas analizy również cech jakościowych poprzez automatyczne przekształcenie ich w odpowiedni sposób.

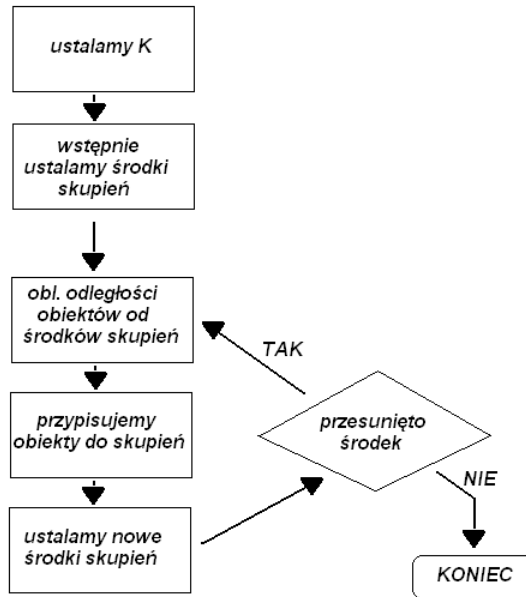
Algorytm ten polega na przenoszeniu obiektów ze skupienia do skupienia w celu zminimalizowania zmienności wewnątrz skupień i zmaksymalizowania zmienności między skupieniami.

Zasada działania algorytmu k-średnich jest następująca – kolejno wykonujemy kroki:

1. Ustalamy liczbę skupień, liczba tych skupień oznaczana jest literą k i stąd nazwa tej metody.
2. Ustalamy wstępnie środki skupień.
3. Obliczamy odległości obiektów od środków skupień.
4. Przypisujemy obiekty do skupień – dla danego obiektu porównujemy odległości do wszystkich środków skupień (obliczone w punkcie 3) i przypisujemy go do tego skupienia, do którego środka ma najbliżej.
5. Ustalamy nowe środki skupień – najczęściej przyjmuje się, że jest to punkt, którego współrzędnymi są średnie arytmetyczne współrzędnych obiektów, które na danym etapie działania algorytmu należą do danego skupienia.
6. Jeśli w punkcie 5 przesunęliśmy środki skupień, to powtarzamy kroki 3, 4, 5, natomiast jeśli nie, to algorytm zatrzymuje się, a za ostateczną segmentację przyjmujemy bieżący podział.

Ilustracja tej procedury znajduje się także na diagramie na rys. 2 (poniżej).

Jak zaznaczono powyżej, przed przystąpieniem do analizy należy określić liczbę skupień, na którą chcemy dzielić interesujące nas obiekty. Aby ustalić optymalną liczbę skupień można skorzystać z szeregu metod ich wyznaczania (są one przedstawione np. w [2]). Metody te opiszemy w dalszej części artykułu.



Rys. 2. Algorytm k-średnich

Dla jakości uzyskanych wyników duży wpływ ma właśnie etap ustalania parametrów algorytmu, czyli kroki 1, 2 oraz określenie, w jaki sposób będzie obliczana odległość między obiektami.

W przypadku zmiennych ilościowych najczęściej stosuje się odległość euklidesową. Odległość euklidesowa, tak jak inne podobne do niej miary odległości, ma jednak pewną wadę, może silnie podlegać wpływowi jednej ze zmiennych, mianowicie tej, której zakres wartości jest największy. Jeśli wartości tej zmiennej są znacznie większe od wartości innych zmiennych, wtedy o różnicy bądź podobieństwie między obserwacjami będzie, w dużej mierze, decydowała tylko ta jedna zmienna (wynika to wprost z formuły, za pomocą której wyliczamy odległość euklidesową). Może to mieć miejsce na przykład, gdy zmienne wyrażone są w różnych jednostkach lub reprezentują różny rząd wielkości.

Aby zapobiec takiej sytuacji, stosuje się normalizację, czyli wartości każdej ze zmiennych (X_j) przekształca w następujący sposób:

$$X'_j = \frac{X_j - \text{Min}(X_j)}{\text{Max}(X_j) - \text{Min}(X_j)},$$

gdzie $\text{Min}(X_j)$, $\text{Max}(X_j)$ oznaczają odpowiednio najmniejszą i największą wartość zmiennej X_j . Po takim zabiegu wszystkie zmienne przyjmują wartości z tego samego przedziału – $[0,1]$. W niektórych przypadkach odchodzi się jednak od procedury



normalizacji, szczególnie w sytuacji, gdy zmienne mają takie same zakresy wartości. Normalizacja może wtedy usunąć różnice między zmiennymi, podczas gdy mogą one nieść ważne informacje (np. jedna ze zmiennych może zawsze przyjmować tylko wartości ujemne, mimo że wartości dodatnie też są dla niej dozwolone, przypominamy, że po normalizacji wszystkie zmienne mają wartości z przedziału $[0,1]$). Z sytuacją taką będziemy mieć do czynienia w naszym przykładzie.

W praktyce oczywiście nie ma jednej uniwersalnej metodyki, jednego uniwersalnego zestawu parametrów (liczby skupień, metody wyznaczania wstępnych centrów skupień, liczby iteracji, sposobu mierzenia podobieństwa między obserwacjami) dającego najlepsze rezultaty dla każdego typu danych. W dalszej części omówimy niektóre aspekty ustalania tych parametrów w odniesieniu do konkretnej analizy.

Prezentacja rozwiązywanego problemu

W ostatnich latach nastąpił ogromny przyrost danych pochodzących z eksperymentów medycznych i genetycznych. Spowodowane to jest postępowaniem w poznawaniu ludzkiego genomu (*Human Genome Project*) oraz technologią mikromacierzy DNA. Mikromacierze umożliwiają badanie w jednym eksperymencie wielu genów - ocenę, które z nich są czynne, a które wyłączone, i jaki jest poziom ich ekspresji. Pozwala to badać mechanizmy regulacyjne żywej komórki, jednak wymaga specjalnych narzędzi do analizy tak dużej liczby danych.

W naszym przykładzie będziemy właśnie analizować taki zbiór danych, gdzie zmiennymi są poziomy ekspresji genów. Do analizy takich danych wykorzystamy opisaną wcześniej analizę skupień.

Zwykle pierwszym etapem analizy jest wstępne zbadanie danych oraz określenie celów analizy. W naszym przypadku podczas badań pobrano próbki DNA od 64 różnych pacjentów z chorobą nowotworową. Dla każdej próbki zbadano ekspresję wybranych 6830 genów. Dane zestawione są w macierzy, w której każdy wiersz reprezentuje próbkę (podaje poziomy ekspresji genów dla danej próbki), natomiast w kolumnach mamy ekspresję poszczególnych genów. Ekspresja każdego genu charakteryzowana jest przez liczbę rzeczywistą mierzącą poziom kwasu mRNA obecnego w danym genie. Będziemy rozpatrywać związki między wierszami macierzy reprezentującej poziomy ekspresji poszczególnych genów.

Każda z próbek ma dodatkowo etykietę mówiącą, z jakiej części organizmu została pobrana. Nie będziemy używać tych etykiet podczas naszej analizy, dopiero na koniec porównamy, czy próbki nowotworów tego samego rodzaju trafiły do tych samych skupień. Oczywiście nie jest to kryterium poprawności analizy, ponieważ nie mamy żadnych przesłanek, aby twierdzić, że próbki pobrane z tych samych tkanek mają tę samą ekspresję genów, a z różnych części - różną, aczkolwiek wydaje się, że tak powinno być.

Warto jeszcze raz podkreślić, że wszystkie zmienne w naszym przykładzie mają wartości w zbiorze liczb rzeczywistych i dodatkowo mają taki sam potencjalny zakres wartości,



dlatego też przed przystąpieniem do analizy nie będziemy wykonywać normalizacji zmiennych.

	1 Zmn1	2 Zmn2	3 Zmn3	4 Zmn4	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9
1	0,3	1,18	0,55	1,14	-0,265	-0,07	0,35	-0,315	
2	0,679961	1,289961	0,169961	0,379961	0,464961	0,579961	0,699961	0,724961	-0,0
3	0,94	-0,04	-0,17	-0,04	-0,605	0	0,09	0,645	
4	0,28	-0,31	0,68	-0,81	0,625	-1,36777900E-17	0,17	0,245	
5	0,485	-0,465	0,395	0,905	0,2	-0,005	0,085	0,11	
6	0,31	-0,03	-0,1	-0,46	-0,205	-0,54	-0,64	-0,585	
7	-0,83	0	0,13	-1,63	0,075	-0,36	0,1	0,155	
8	-0,19	-0,87	-0,45	0,08	0,005	0,35	-0,04	-0,265	
9	0,46	0	1,15	-1,4	-0,005	-0,7	-0,92	-0,515	
10	0,76	1,49	0,28	0,1	-0,525	0,36	0,6	0,175	
11	0,27	0,63	-0,36	-1,04	0,015	-0,04	0,58	0,425	
12	-0,45	-0,06	0,15	-0,61	-0,395	0,15	0,94	-0,155	
13	-0,03	-1,12	-0,05	0	-0,285	-0,25	0,16	0,085	
14	0,71	0	0,16	-0,77	0,045	-0,16	0,32	-0,025	
15	-0,36	-1,42	-0,03	-2,28	0,135	-0,32	-0,04	0,035	
16	-0,21	-1,95	-0,7	-1,65	-0,075	0,06	-0,06	0,015	
17	-0,5	-0,52	-0,66	-2,61	0,225	-0,05	-0,39	-0,095	
18	-1,06	-2,19	-0,13	0	-0,485	-0,43	-0,09	0,045	
19	0,15	-0,45	-0,32	-1,61	-0,095	-0,08	0	0,225	-1,62969
20	-0,29	0	0,05	0,73	0,385	0,39	0	-0,065	
21	-0,2	0,74	0,08	0,76	-0,105	-0,08	-0,31	-0,465	
22	0,43	0,5	-0,73	0,6	-0,635	-0,43	-0,22	0,015	

Rys. 3. Arkusz danych *ncidata.sta*

Na etapie zapoznawania się z danymi warto również sprawdzić, czy nie ma obserwacji odstających, ponieważ wówczas podczas analizy skupień prawdopodobnie zostanie utworzone jedno skupienie zawierające tę obserwację odstającą, a wszystkie pozostałe przypadki mogą zostać zakwalifikowane do jednego skupienia.

Co będzie celem naszej analizy? Przede wszystkim chcielibyśmy się dowiedzieć, czy istnieją jakieś różnice pomiędzy nowotworami ze względu na poziom ekspresji genów, a jeśli tak, to czy można je jakoś scharakteryzować. Zależałoby nam również na tym, aby wyodrębnić jednorodne grupy nowotworów podobnych. Będziemy się starali pogrupować próbki, biorąc pod uwagę tylko poziomy ekspresji genów. Ponieważ mamy do dyspozycji bardzo dużo zmiennych, aż 6830, interesowałoby nas więc także to, które z nich są najistotniejsze, które mają największy wpływ na uzyskany podział. Spróbujemy znaleźć te zmienne, po czym sprawdzimy, czy dla wybranego podzbioru zmiennych mamy podobne wyniki, czyli ocenimy, na ile dobry jest wybrany podzbiór predyktorów. Do powyższych analiz użyjemy metodę *k*-średnich.

Analiza w środowisku *STATISTICA Data Miner*

Do przeprowadzenia analizy możemy wybrać *Grupowanie metodą k-średnich* z modułu Wielowymiarowe techniki eksploracyjne lub *Analizę skupień uogólnioną metodą k-średnich* z modułu Uogólniona analiza skupień. Podczas wykonywania analizy

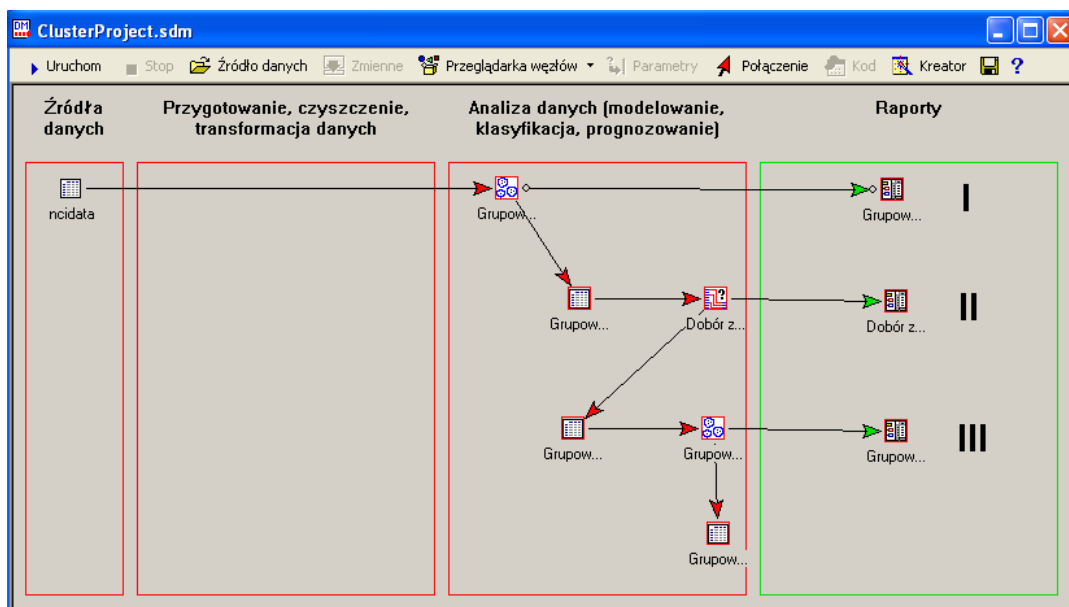


uogólnioną metodą k -średnich we wstępnej fazie wykonywana jest normalizacja zmiennych, toteż z powodów opisanych wcześniej wybierzemy zwykłe *Grupowanie metodą k -średnich*.

Po wykonaniu segmentacji postaramy się znaleźć te zmienne, które miały największy wpływ na przeprowadzony podział. Wybierzemy 50 najlepszych predyktorów. Do tego celu użyjemy modułu *Dobór zmiennych i analiza przyczyn*, a następnie zobaczymy, jaki podział uzyskamy, używając tylko tych wybranych zmiennych.

Tak więc plan naszej analizy to: analiza skupień na całości danych (I), potem wybór najlepszych predyktorów (II), a następnie analiza skupień dla nich (III).

Wszystkie analizy przeprowadzimy w przestrzeni roboczej *STATISTICA Data Miner*, dzięki czemu w jednym projekcie otrzymamy wszystkie wyniki. Dodatkowo widzimy i możemy kontrolować przebieg wszystkich analiz nawet w bardzo złożonym projekcie, wygodnie dodawać nowe metody oraz zmieniać dane wejściowe.



Rys. 4. Przestrzeń robocza programu *STATISTICA Data Miner*

Aby rozpocząć analizę, wybieramy opcję *Data Miner–Wszystkie procedury* z menu *Statystyka–Data-Mining*. Na ekranie pojawi się przestrzeń robocza programu *STATISTICA Data Miner*. Za pomocą przycisku *Źródło danych* wybieramy dane wejściowe, a za pomocą *Przeglądarki węzłów* wybieramy odpowiednie procedury.

Każda procedura przetwarzająca dane reprezentowana jest przez ikonę (tzw. węzeł). Przepływ danych obrazują strzałki łączące poszczególne węzły. Niektóre analizy jako wyniki zwracają, prócz skoroszytu wyników, także arkusze danych, które można dalej przekształcać. Węzły zaprojektowane są tak, aby dane wpływające z jednego z węzłów



mogły stanowić wejście dla innych węzłów. Zapewnia to możliwość składania projektu analizy z poszczególnych elementów.

Budując projekt, w przeglądarce węzłów zaznaczamy odpowiedni węzeł i wstawiamy go do przestrzeni roboczej (przycisk *Wstaw*), łącząc go z odpowiednim arkuszem danych. Na koniec klikamy przycisk *Uruchom* na pasku narzędzi przestrzeni roboczej, aby uruchomić projekt.

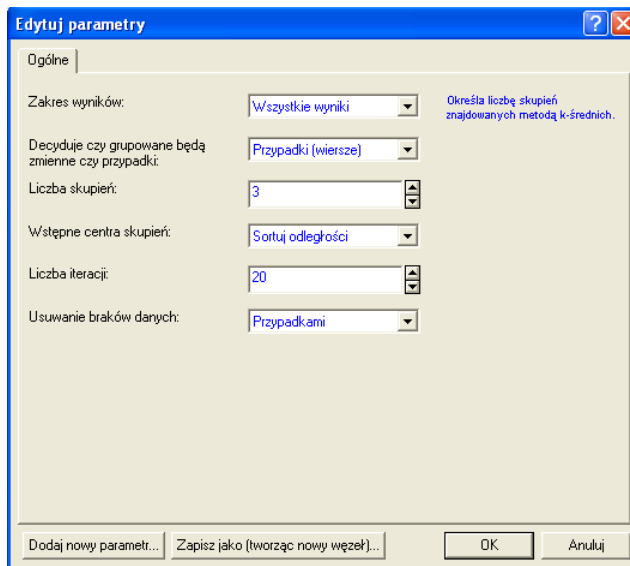
Dodatkową zaletą programu *STATISTICA Data Miner* jest to, że każdy węzeł można zmodyfikować. Klikając na odpowiednim węźle, można podglądać i edytować jego kod napisany w języku *STATISTICA Visual Basic* (jest to język Visual Basic wzbogacony o procedury statystyczne). W naszym przykładzie niektóre węzły również zostały nieznacznie zmodyfikowane.

Grupowanie metodą k-średnich

Aby przeprowadzić zaplanowaną analizę, do przestrzeni roboczej wstawiamy plik danych *ncidata.sta*. Następnie z *Przeglądarki węzłów* wybieramy węzeł: *Grupowanie metodą k-średnich*. Wykonując analizę metodą k-średnich, musimy ustalić kilka ważnych parametrów tej analizy, jak zostało to zaznaczone w opisie tego algorytmu. Ustalamy zmienne, których będziemy używać do analizy, ustalamy liczbę skupień, wstępne ich centra, sposób mierzenia odległości między grupowanymi obiektami oraz liczbę iteracji, jaką wykona algorytm.

Wybór zmiennych

Analizę rozpoczniemy od zbudowania modelu przy użyciu wszystkich zmiennych.



Rys. 5. Karta wyboru parametrów algorytmu w *Grupowaniu metodą k-średnich*



W oknie wyboru zmiennych, w naszym arkuszu danych podłączonym do węzła *Grupowanie metodą k-średnich*, zaznaczamy wszystkie 6830 zmienne.

Przechodzimy na kartę *Edytuj parametry* procedury k-średnich (rys. 5). Wybieramy opcję grupowania danych przypadkami i przechodzimy do edycji pozostałych parametrów:

Wybór liczby skupień

Wybór liczby skupień może być dokonany na wiele sposobów. Jedną z metod jest po prostu umowne ustalenie liczby skupień i ewentualna późniejsza zmiana tej liczby, w taki sposób, aby otrzymać lepsze wyniki. Wstępne ustalenie liczby skupień może być oparte na wynikach innych analiz.

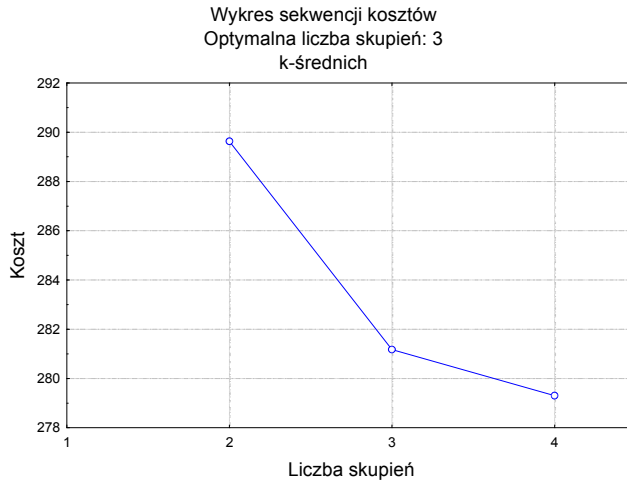
Metodą proponowaną przez Guidiciego [2] jest przeprowadzenie wstępnej analizy za pomocą metody hierarchicznej, oszacowanie za jej pomocą liczby skupień, a następnie dla tak wybranej liczby skupień wykonanie już analizy metodą niehierarchiczną, czyli na przykład właśnie metodą k-średnich. Metody hierarchiczne są jednak ograniczone. Przy zbyt dużej liczbie danych wstępna ocena liczby skupień wykonywana jest na podzbiórce danych, co jest niewątpliwie pewną niedogodnością.

Alternatywnym podejściem w tej sytuacji jest przeprowadzenie najpierw analizy niehierarchicznej i stworzenie dużej liczby skupień, a potem dalsze grupowanie za pomocą metody hierarchicznej, przy uwzględnieniu odległości i liczebności skupień. W tym przypadku, wstępnie przeprowadzone grupowanie metodą k-średnich ma na celu zredukowanie liczby danych (w drugiej części analizy grupujemy już tylko skupienia, nie biorąc pod uwagę ich poszczególnych elementów).

Podczas korzystania z modułu *Uogólniona analiza skupień metodą k-średnich* można skorzystać ze sprawdzianu krzyżowego do oceny liczby skupień. Wydaje się to być najlepsza metoda wyboru liczby skupień. Nie ma w tym przypadku ingerencji w analizę (brak założeń *a priori* o liczbie skupień), problemów z wyborem podzbiórki danych, ani w żaden sposób nie jesteśmy także ograniczeni, co jest istotne, liczbą danych.

Algorytm ten dzieli zbiór wejściowy kolejno na coraz większą liczbę segmentów, a następnie sprawdza, 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.

Wyniki sprawdzianu krzyżowego ilustrowane są na tak zwanym wykresie osypiska (rys. 6). Analizując wykres, można zauważyć znaczną poprawę precyzji podziału przy zwiększeniu liczby segmentów z dwóch do trzech. Dodając jeszcze jeden segment, uzyskuje się już znacznie mniejszą poprawę precyzji, stąd za optymalną liczbę segmentów należy uznać trzy. Program *STATISTICA* automatycznie określa najbardziej odpowiednią liczbę skupień.



Rys. 6. Przykładowy wykres osypiska

W naszym przykładzie, tak jak to jest również w pracy [1], będziemy dzielić interesujące nas obserwacje na 3 skupienia.

Wybór wstępnych centrów skupień

Do wyboru mamy trzy możliwości:

- ◆ Wybierz obserwacje tak, by zmaksymalizować odległości skupień.
- ◆ Sortuj odległości i weź obserwacje przy stałym interwale.
- ◆ Wybierz pierwszych N (liczba skupień) obserwacji.

Wybieramy domyślną opcję wyznaczenia jako początkowych centrów skupień obiektów przy stałych interwałach. W *Analizie skupień uogólnioną metodą k-średnich* mamy dodatkowo *Losowy wybór N obserwacji*. Jedną z zalecanych metod jest sprawdzenie i porównanie wyników z kilkakrotnie przeprowadzonej analizy, gdy wstępne centra wybierane były w sposób losowy, i wybranie najlepszego modelu. Zapobiega to trafieniu w lokalne minimum, przy minimalizowaniu wewnętrznej wariancji w skupieniach, podczas procesu doboru obserwacji do skupień.

Wybór odległości

Domyślnie w *Grupowaniu metodą k-średnich* mamy przeskalowaną odległość euklidesową. Odległość między dwoma obiektami lub centrami skupień X_i i X_j obliczana jest na podstawie wzoru

$$D(i, j) = \sqrt{\frac{1}{M} \sum_{k=1}^M (X_{ik} - X_{jk})^2},$$

gdzie M to liczba zmiennych (wymiar przestrzeni).

W Analizie skupień uogólnioną metodą k -średnich dodatkowo można wybrać inną odległość. Do wyboru mamy jedną z następujących odległości: odległość euklidesowa, kwadrat odległości euklidesowej, odległość Manhattan, odległość Czebyszewa.

Ile iteracji

Ostatnim parametrem, jaki należy ustalić, jest określenie liczby iteracji wykonanych podczas analizy. Jak zaznaczono powyżej warunkiem zatrzymania algorytmu jest brak przesunięcia obiektów pomiędzy skupieniami. Jeśli jednak algorytm wykona zadaną przez ten parametr liczbę iteracji, to proces analizy zostanie zatrzymany, nawet jeśli powyższy warunek zatrzymania procedury nie zostanie spełniony. Zostawiamy domyślną wartość tego parametru, mianowicie 20.

ID przyp.	Średnia	Standar. Odchylenie	Warianc.
C_1	0,063823	0,493754	0,24379
C_2	0,063823	0,924530	0,85476
C_3	-0,001324	0,408246	0,16667
C_4	-0,310736	1,088766	1,18541
C_5	0,105588	0,468985	0,21995
C_6	-0,023530	0,329884	0,10882
C_7	0,103823	0,422841	0,17879
C_8	0,027646	0,321278	0,10322
C_9	0,012941	0,462420	0,21383
C_10	-0,094393	0,748429	0,56015
C_11	0,143823	0,678062	0,45977
C_12	0,127941	0,674158	0,45449
C_13	-0,018236	0,361678	0,13081
C_14	0,028095	0,386025	0,14902
C_15	0,532647	1,125037	1,26571
C_16	0,871618	2,776962	7,71152
C_17	0,231764	1,230515	1,51417
C_18	0,018382	0,427555	0,18280
C_19	0,115018	0,631358	0,39861
C_20	0,020140	0,455104	0,20713

Rys. 7. Skoroszyt wyników *Grupowania metodą k-średnich*

Po określeniu wszystkich parametrów klikamy przycisk *Uruchom* na pasku narzędzi przeszerzeni roboczej, aby rozpocząć analizę. Wynikiem analizy jest skoroszyt arkuszy (rys. 7) zawierających:

- ♦ elementy każdego skupienia – arkusz pokazujący, do jakich skupień zakwalifikowane zostały poszczególne przypadki,
- ♦ średnie skupień,
- ♦ średnie dla każdego skupienia zestawione na jednym wykresie,
- ♦ statystyki opisowe dla wszystkich skupień – mogą być użyteczne do scharakteryzowania skupień i opisu ich własności,
- ♦ odległości euklidesowe między skupieniami – zestawione w macierzy odległości,

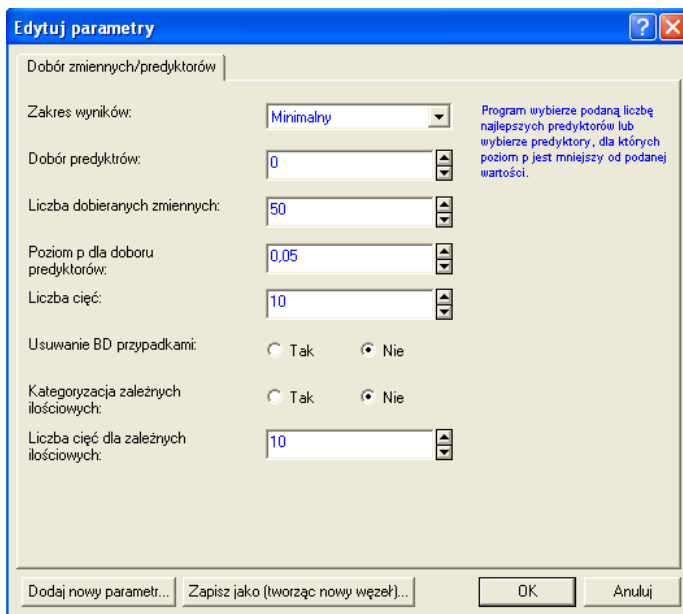


- ◆ analiza wariancji – kolejno dla każdej zmiennej, porównujemy ze sobą jej średnie we wszystkich segmentach, im istotniejsze różnice między średnimi (mniejsza wartość p), tym dana zmienna bardziej różnicuje skupienia.

Analizując otrzymane wyniki, możemy stwierdzić, że otrzymaliśmy podział na trzy skupienia o licznosciach odpowiednio 9, 21, 34. Teraz postaramy się znaleźć te zmienne, które miały największy wpływ na uzyskany podział.

Poszukiwanie najlepszych predyktorów

Wynikiem poprzednio zastosowanego węzła był także arkusz zawierający dane wejściowe oraz dodatkową kolumnę, mianowicie wynikową segmentację. Każdemu przypadkowi została przyporządkowana liczba 1, 2 lub 3 mówiąca, do którego skupienia dany przypadek został zakwalifikowany. Teraz do tego arkusza dołączamy węzeł: *Dobór zmiennych i analiza przyczyn*.

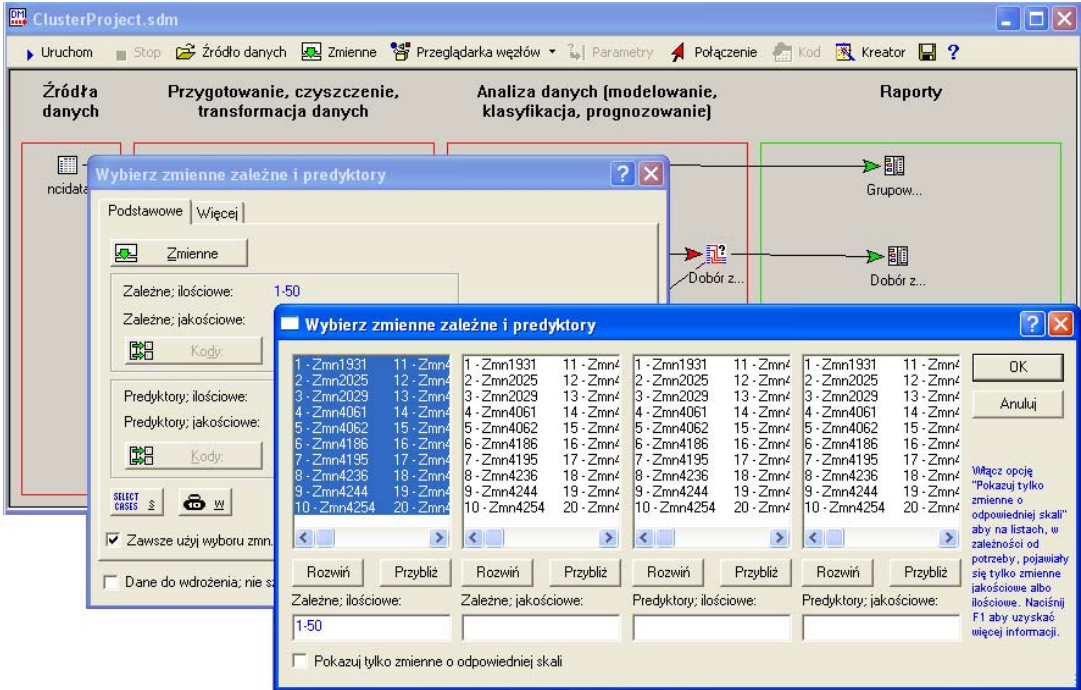


Rys. 8. Karta wyboru parametrów w węźle *Dobór zmiennych i analiza przyczyn*

Podczas analizy wybierane są te zmienne, które wpływają na badaną cechę, w naszym przypadku numer skupienia (1, 2, 3), do którego dany przypadek trafił. Zmienna zależna w naszym przykładzie ma charakter jakościowy, program oblicza więc statystykę χ^2 (chi-kwadrat) oraz wartość p dla każdego predyktora. W przypadku predyktorów ilościowych zakres wartości predyktora, poziom ekspresji poszczególnych genów, dzielony jest na k przedziałów (domyślnie 10). Gdyby występowały dodatkowo predyktory jakościowe, nie byłyby one przekształcane w żaden sposób. Na karcie doboru parametrów tego węzła ustalamy, prócz liczby cięć, ile zmiennych ma być wybranych (my wybieramy 50) oraz



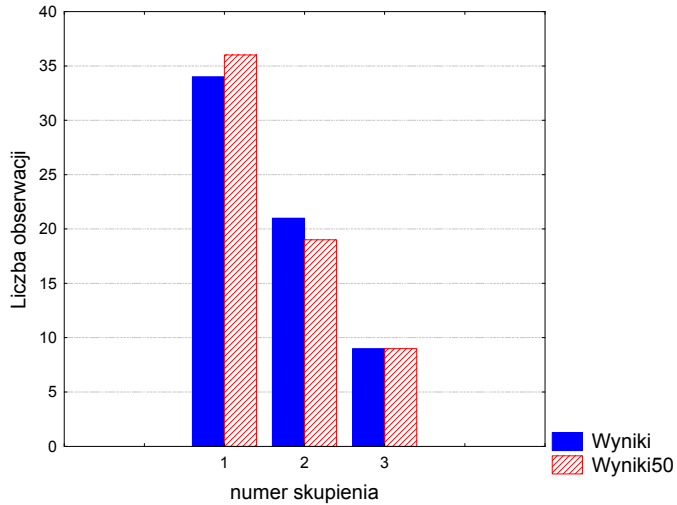
poziom p dla doboru tych zmiennych. Wynikiem jest arkusz, w którym wybranych jest już 50 najlepszych predyktorów.



Rys. 9. Okno wyboru zmiennych dla wynikowego arkusza danych węzła *Dobór zmiennych i analiza przyczyn*

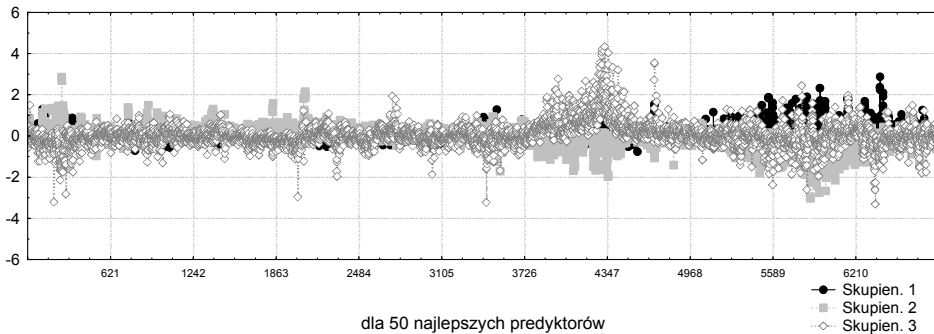
Taki arkusz z wybranymi zmiennymi jest gotowy do dalszych analiz. Przeprowadzamy na nim analizę skupień metodą k-średnich, z parametrami ustawionymi jak poprzednio. Otrzymujemy w ten sposób segmentację opartą na 50 zmiennych. Otrzymaliśmy skupienia o licznosciach 9, 19, 36. Przy porównaniu elementów skupień okazuje się, że tylko dwie obserwacje zostały przydzielone do innych skupień w porównaniu z analizą na całości danych (rys. 10).

Zatem nowy podział na skupienia niemalże pokrywa się z tym uzyskanym, gdy bierzemy pod uwagę wszystkie zmienne. Wyniki segmentacji na ogół weryfikuje się poprzez porównywanie wartości średnich wartości cech w skupieniach. Na rys. 11 w górnej części ze względu na liczbę zmiennych trudno jest dostrzec jakieś prawidłowości, natomiast na dole widać, że średnie w wydzielonych skupieniach różnią się między sobą znacznie. Z wykresu średnich każdego skupienia dla wszystkich zmiennych trudno wyciągnąć jakieś wnioski, natomiast gdy rozpatrujemy tylko 50 zmiennych, można już zauważyć, w jaki sposób poziom ekspresji pewnego genu determinuje przydział do danego skupienia.

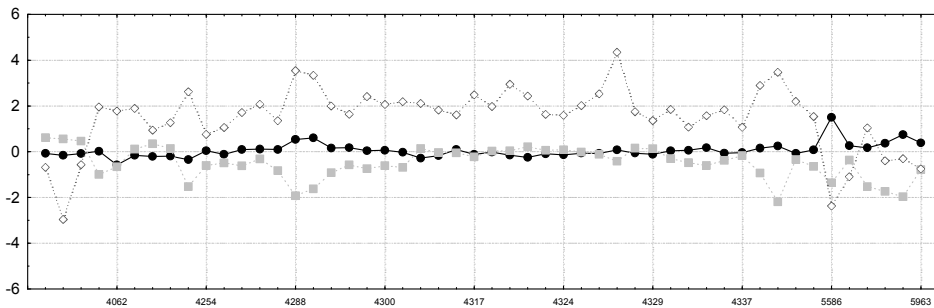


Rys. 10. Wykres liczności każdego skupienia w przypadku analizy dla wszystkich zmiennych (po lewej) i dla 50 najlepszych predyktorów

Wykres średnich każdego skupienia dla wszystkich zmiennych



dla 50 najlepszych predyktorów



Rys. 11 Wykres średnich każdego skupienia dla wszystkich zmiennych (na górze) i dla 50 najlepszych predyktorów

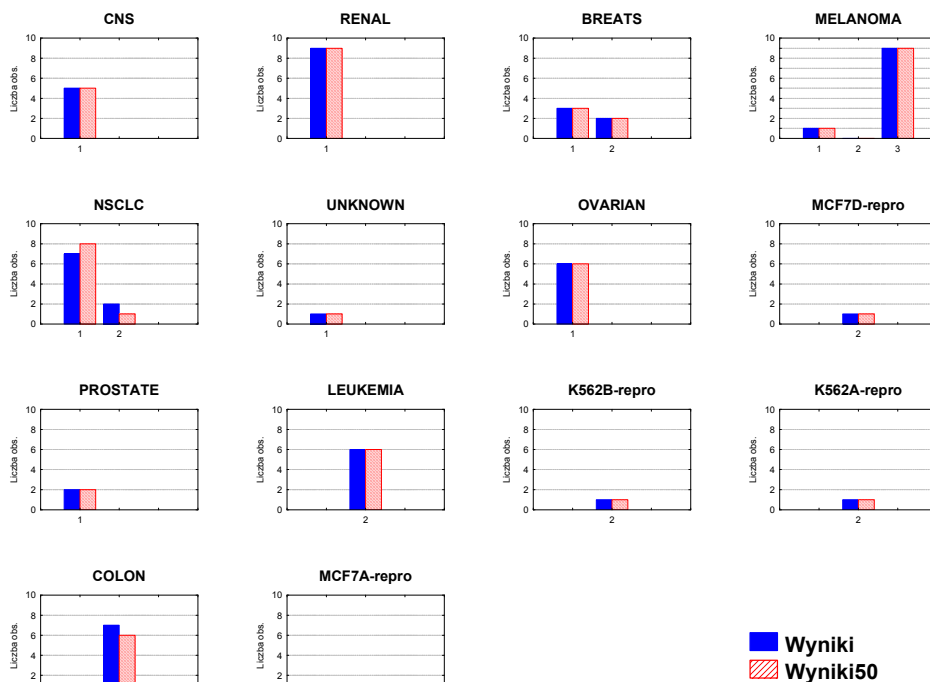


Omówienie wyników

Hipotezę sformułowaną we wcześniejszej części pracy, że wydzielone skupienia będą pokrywać się z podziałem na rodzaje nowotworów, których to próbki analizowaliśmy, wydaje się być prawdziwa, ponieważ prawie wszystkie próbki nowotworów tego samego rodzaju znalazły się w tych samych skupieniach.

Etykiety	Wyniki 1	Wyniki 2	Wyniki 3	Wiersz Razem
CNS	5	0	0	5
RENAL	9	0	0	9
BREAST	3	2	0	5
NSCLC	7	2	0	9
UNKNOWN	1	0	0	1
OVARIAN	6	0	0	6
MELANOMA	1	0	9	10
PROSTATE	2	0	0	2
LEUKEMIA	0	6	0	6
K562B-repro	0	1	0	1
K562A-repro	0	1	0	1
COLON	0	7	0	7
MCF7A-repro	0	1	0	1
MCF7D-repro	0	1	0	1
Ogół grp	34	21	9	64

Rys. 12. Porównanie wyników segmentacji z rodzajem nowotworu (dla modelu wykorzystującego wszystkie zmienne)



Rys. 13. Zestawianie wyników dla wszystkich i 50 najlepszych predyktorów



Jedynie pojedyncze przypadki nowotworów BREAST, MELANOMA, NSCLC zostały rozrzucone po 2 skupieniach (por. rys. 12 oraz rys. 13).

Jakość wyboru 50 najlepszych predyktorów została zweryfikowana kolejną analizą skupień. Okazało się, że ograniczenie liczby zmiennych z 6830 do 50 nie zaburza wcześniejszej segmentacji. Grupy utworzone podczas analizy skupień na zredukowanych danych mają, poza dwoma wyjątkami, dokładnie taki sam skład.

Wynikami naszej analizy są więc jednorodne skupienia, dzielące wejściowe dane ze względu na poziom ekspresji genów. Widzimy, że podział ten ma związek z rodzajem nowotworu, w kolejnym kroku badacz może charakteryzować poszczególne skupienia ze względu na różne cechy (np. diagnozę, leczenie). Przyporządkowywanie do skupień nowych przypadków może odbywać się już na podstawie poziomów ekspresji zaledwie 50 genów.

Oczywiście te same analizy można przeprowadzić również innymi metodami, na przykład za pomocą drzew hierarchicznych, analizy skupień metodą EM lub sieci Kohonena. Uzyskane wyniki mogą pomóc w zbudowaniu najlepszego modelu.

Podsumowanie

Analiza skupień wydaje się być nieodzownym narzędziem wszędzie tam, gdzie mamy do czynienia z ogromnymi ilościami danych, w których nie widać jakiegokolwiek struktury, a analizowanie pojedynczych przypadków traci sens. Oczywiście może to mieć miejsce jak w opisanym przykładzie w medycynie, ale także w innych dziedzinach badań. Analizę skupień z powodzeniem wykorzystuje się na przykład w wyodrębnianiu segmentów rynku w badaniach marketingowych, wzorców pogody w meteorologii, ścieżek zakupów w analizie zachowań klientów, wzorców zachowań użytkowników serwisów internetowych itd.

Literatura

1. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, Springer 2002.
2. Guidici P., *Applied Data Mining - Statistical Methods for Business and Industry*, John Wiley & Sons, Inc, 2003.