Jerzy Kucharczyk

Algorytmy analizy skupień w języku ALGOL 60,

Państwowe Wydawnictwo Naukowe, Warszawa 1982, str. 275, nakład 3500 + 250 egz., cena 120 zł, ISBN 83-01-02360-0.

W wielu dziedzinach badań naukowych pojawia się problem wydzielenia z danego zbioru obiektów (jednostek, elementów) mniejszych jednorodnych grup, zwanych też klasami lub skupieniami. Obiekty należące do tych samych grup muszą być w pewnym sensie do siebie podobne. Proces grupowania obiektów odbywa się na podstawie pewnych reguł sformułowanych w odniesieniu do wartości liczbowych charakteryzujących te obiekty. Problemy grupowania, nazywane często niezbyt trafnie metodami analizy skupień, są przedmiotem badań matematyczno-statystycznych, wiążąc się najmocniej z tzw. analizą wielowymiarową.

W ostatnich piętnastu latach, dzięki ogromnemu rozwojowi technik komputerowych, nastąpił gwałtowny rozwój metod analizy skupień, co uwidoczniło się pojawieniem dużej liczby opracowań naukowych z tego zakresu. Do tej grupy opracowań należy książka Jerzego Kucharczyka "Algorytmy analizy skupień w języku ALGOL 60", która jest pierwszą tego typu publikacją na naszym rynku wydawniczym.

Książka ta składa się z dwóch części. Pierwsza z nich poświęcona jest opisowi 63 procedur i funkcji niestandardowych, z których 32 procedury pozwalają wykryć istniejące skupienia, natomiast pozostałe stanowią zbiór algorytmów użytecznych w obliczeniach wstępnych oraz przy interpretacji wyników. Tę część książki tworzą następujące rozdziały: I. Procedury pomocnicze, II. Hierarchiczne metody aglomeracyjne, III. Hierarchiczne metody podziału, IV. Interpretacja wyników metod hierarchicznych, V. Metody oparte na uporządkowaniu obiektów, VI. Metody niehierarchiczne, korzystające z macierzy danych, VII. Metody niehierarchiczne, korzystające z macierzy odległości, VIII. Metody

niehierarchiczne, korzystające z pamięci pomocniczej, IX. Interpretacja wyników i porównanie metod, X. Obliczenia wstępne. Część drugą książki stanowią napisane w języku ALGOL 60 teksty procedur i funkcji omawianych w części pierwszej.

Bardziej wnikliwa analiza przedstawionych rozdziałów pozwala stwierdzić, że książka zawiera przegląd najczęściej stosowanych w połowie lat siedemdziesiątych metod grupowania rozłącznego. Zważywszy, że maszynopis książki był gotowy w 1979 roku, uważam, że Autor dokonał trafnego wyboru metod.

Chciałbym teraz podać kilka uwag ogólnych, które nasunęły mi się przy czytaniu tej książki. Pierwsza z nich dotyczy terminu "analiza skupień", który, jak wspomniałem, jest terminem niezbyt trafnym. Bierze się to stąd, że dokonując bezpośredniego tłumaczenia terminu angielskiego "cluster analysis" na termin "analiza skupień" uzyskuje się niezamierzony efekt sugerujący, że ustalony układ skupień poddawany jest analizie, podczas gdy właściwym celem metod grupowania jest wykrywanie samych skupień. W świetle powyższego wydaje się, że trafniejszym odpowiednikiem angielskiego "cluster analysis" byłby termin "poszukiwanie skupień". Te terminologiczne rozważania nasuwają się przy czytaniu Wstępu, gdzie zostało podanych 7 przykładów służących do testowania procedur. W każdym z nich w "naturalny" sposób wyróżniono pewną liczbę skupień. Fakt ten może wprowadzić czytelnika w błąd sugerując, że przedmiotem rozważań jest analiza skupień "naturalnych", a nie ich tworzenie. Należy tu wyjaśnić, co zresztą Autor uczynił, że celem wprowadzenia przykładów, mających "naturalne" skupienia, jest pomoc użytkownikowi w kontroli poprawności działania programów.

Druga z uwag dotyczy sposobu prezentacji algorytmów. Podanie samych opisów procedur dla wybranych metod, bez ich bliższego omówienia, sprawia, że książka ta może być właściwie wykorzystana jedynie przez tych czytelników, którzy metody poszukiwania skupień dobrze znają. Opinię tę postaram się bliżej sprecyzować. W rozdziałach (II, III, IV) poświęconych wykrywaniu skupień. Autor podaje opisy kilku procedur opracowanych o taką samą lub zbliżoną strategię grupowania. Użytkownik, znający słabo zagadnienie skupiania, stanie przed niemałym dylematem wyboru, nie uzyskując najmniejszej wskazówki. Sądzę, że podanie czasów testowania poszczególnych procedur zapisanych w dowolnej, lecz jednakowej realizacji ALGOL-u dla przykładów testowanych przybliżyłoby czytelnikowi wybór odpowiedniej metody. W każdym razie pozwoliłoby na wybranie metody taniej, co w zagadnieniach skupiania jest sprawą dość istotną.

Pewną niekonsekwencją książki jest umieszczenie w spisie publikacji artykułu Fishera [2], który jednakże nie został w książce wykorzystany. A szkoda, bo praca ta zasługuje na szczególną uwagę, gdyż wskazuje, jak metody programowania matematycznego należy wykorzystać w problemie poszukiwania skupień. Fakt ten stał się podstawą rozwoju, w pierwszej połowie lat siedemdziesiątych, bardzo szybkich metod grupowania. Całkowite pominięcie tych metod poważnie książkę zuboża. Dla użytkowników tworzących własne biblioteki procedur może w związku z tym być pomocna następująca informacja. Otóż, mocniejszym kryterium podziału jest badanie zmienności wewnątrzskupieniowej (patrz Harabasz i Wiśniewski [5]) niż zaproponowane w rozdzia-

le III badanie odległości między obiektami skupień. Warto zwrócić tu uwagę, że Autor odwołuje się w tym miejscu do pracy Huberta [6], chociaż są znane wcześniejsze algorytmy wykorzystujące metody programowania matematycznego, które pozwalają wykryć istniejące skupienia przez badanie odległości między obiektami skupień (patrz np. Rao [7]).

Innym niedostatkiem książki jest niedość wyraźne wydzielenie procedur pozwalających skupiać obiekty jednocechowe (nb. wszystkie testowane przykłady podano dla przestrzeni R^2). Problem ten pojawia się bardzo często w badaniach statystycznych, gdzie zachodzi potrzeba grupowania średnich obiektowych jako uzupełnienie wnioskowania z jednowymiarowej analizy wariancji. Dąży się przy tym do wyróżnienia grup rozłącznych, wewnętrznie jednorodnych. Taki cel nie jest na ogół osiągany za pomocą różnych znanych w literaturze metod porównań wielokrotnych. Te ostatnie prowadzą bowiem najczęściej do wydzielania grup wzajemnie na siebie zachodzących, co zwykle przysparza eksperymentatorowi sporo trudności interpretacyjnych.

Dla skupiania obiektów jednocechowych mogę polecić bardzo efektowny algorytm skupiania dynamicznego, przedstawiony przez Harabasza i Wiśniewskiego [4]. Chcąc zilustrować szybkość działania tego algorytmu, wystarczy powiedzieć, że dla pogrupowania 55 obiektów na 2, 3, ..., 54 skupienia tak, aby wewnątrzskupieniowa suma kwadratów osiągnęła minimum, potrzeba około 200 s przy realizacji obliczeń na emc ODRA-1204.

Na koniec jeszcze jedna uwaga, tym razem o wyborze najlepszej liczby skupień. W rozdziale IX podano wskaźnik jakości podziału obiektów na skupienia, oparty na pewnej a priori wy-

branej "krytyoznej" odległości. Jest to postępowanie subiektywne, czemu można zaradzić, określając tzw. wskaźnik informacyjny dla najlepszej liczby grup (patrz Caliński i Harabasz [1], Harabasz i Karoński [3]).

Mimo tych kilku krytycznych uwag, książkę Jerzego Kucharczyka należy ocenić jako pozycję bardzo wartościową. Dostarcza ona szerokim kręgom użytkowników technik komputerowych wielu procedur umożliwiających poszukiwanie skupień, co stanowi podstawę ich dalszego analizowania. Należy żywić nadzieję, że książka ta stanie się punktem wyjścia dalszego rozwoju biblioteki procedur dla wykrywania i analizowania skupień.

PRACE CYTOWANE

[1] T. C a l i ń s k i, J. H a r a b a s z, A dendrite method for cluster analysis, Communications in Statistics 3 (1974), 1-27.

[2] W.D. F i s h e r, On grouping for maximum homogenity, J. Amer. Statist. Assoc. 53(1958), 789-798.

[3] J.S. H a r a b a s z, M. K a r o ń s k i, Dendrytowa metoda analizy skupień, Roczniki Akademii Rolniczej w Poznaniu, ABS-57(1977), 135-148.

[4] J.S. H a r a b a s z, P. W i ś n i e w s k i, Grupowanie obiektów jednocechowych za pomocą programowania dynamicznego, Roczniki Akademii Rolniczej w Poznaniu, ABS-109 (1983).

[5] J.S. H a r a b a s z, P. W i ś n i e w s k i, Generowanie programu liniowego dla problemu grupowania, Roczniki Akademii Rolniczej w Poznaniu, ABS-108(1983).

[6] L. H u b e r t, Monotone invariant clustering procedures, Psychometrika 38 (1973), 47-62.

[7] M.R. R a o, Cluster analysis and mathematical programming, A. Amer. Statist. Assoc. 66(1971), 622-626.

PIOTR WIŚNIEWSKI