О. Л. ТОЦЬКА

Волинський державний університет імені Лесі Українки

АВТОМАТИЗАЦІЯ МЕТОДУ К-СЕРЕДНІХ КЛАСТЕРНОГО АНАЛІЗУ ЗА ДОПОМОГОЮ ПРОГРАМНОГО ПАКЕТА STATSOFT STATISTICA 6.0

Кластерний аналіз використовується для розділення множини вхідних даних на однорідні групи. Розрізняють такі три методи кластерного аналізу, як ієрархічний, К-середніх, двовходове об'єднання. Основними завданнями методу К-середніх є:

1) мінімізувати змінність всередині кластерів;

2) максимізувати змінність між кластерами.

При його застосуванні потрібно наперед задавати кількість кластерів, яку бажаємо отримати. Оптимальну їхню кількість можна знайти або за допомогою підбору, або попереднім застосуванням ієрархічного методу. При попередній ієрархічній класифікації потрібно переглянути список об'єднання, у якому будуть вказані відстані об'єднання. Оптимальною вважається кількість кластерів, яка дорівнює різниці кількості спостережень і кількості кроків, після якої відстань об'єднання збільшується скачкоподібно.

Для автоматизації методу К-середніх кластерного аналізу за допомогою пакета StatSoft Statistica 6.0 потрібно виконати наступні дії:

1) завантажити програму, створити файл з даними, вибрати метод кластерного аналізу (у нашому випадку – K-means clustering);

2) вказати початкові параметри

на вкладці "Advanced" заповнити вказані нижче рядки і натиснути кнопку OK:

у рядку "Cluster" (Кластер) вказати, що буде класифікуватись:

- Variables (columns) – змінні (колонки),
- Cases (rows) – спостереження (рядки);

у рядку "Number of clusters" ввести кількість кластерів (2 за замовчуванням);

у рядку “Number of iterations” ввести кількість ітерацій (10 за замовчуванням);

у блоці “Initial cluster centers” (Центри початкового кластера) поставити перемикач на одній з трьох опцій:

- Choose observations to maximize initial between-cluster distances – вибрати спостереження для максимізації початкових міжгрупових дистанцій,
- Sort distances and take observations at constant intervals – відсортувати відстані і взяти спостереження в постійних інтервалах,
- Choose the first N (Number of clusters) observations – вибрати перших N (кількість кластерів) спостережень;

у блоці “MD (missing data) deletion” (Вилучення відсутніх даних) поставити перемикач на одній з двох опцій:

- Casewise – мудрий реєстр (видаляє рядки чи стовпці з відсутніми даними),
- Mean substitution – заміна середнім;

у рядку “Batch processing and reporting” (Обробка пакету даних і повідомлення) поставити при потребі прапорець;

3) ознайомитись з результатами

Summary: cluster means & Euclidean distances – відомість: кластер середніх значень і Евклідові відстані,

Analysis of variance – дисперсійний аналіз,

Graph of means – графік середніх значень,

Descriptive statistics for each cluster – описова статистика для кожного кластера,

Members of each cluster & distances – члени кожного кластера і відстані,

Save classifications and distances – зберегти класифікації і відстані;

4) зберегти результати

File (Файл) → Save (Зберегти) → у рядку “Ім’я файла” ввести назву → Зберегти.

Olesya Totska

Volyn' Lesya Ukrayinka state university

AUTOMATION OF K-MEANS METHOD OF THE CLUSTER ANALYSIS BY MEANS PROGRAM PACKAGE STATSOFT STATISTICA 6.0

In the theses of conference the process of realization of K-means method of the cluster analysis is described. Automation is conducted by means program package StatSoft Statistica 6.0. The method of determination of optimum number of clusters is