УДК 519.688

ЗАСТОСУВАННЯ КЛАСТЕРИЗАЦІЇ ДЛЯ ОБРОБКИ РЕЗУЛЬТАТІВ СПОСТЕРЕЖЕННЯ СТАНУ НАВКОЛИШНЬОГО СЕРЕДОВИЩА

[1]Гордійко Н.О., [2]Томашевська Т.В.
[1]*Фізико-технічний інститут КПІ ім. Ігоря Сікорського,*
[2]*Національна академія статистики, обліку та аудиту*
E-mail: natalygor22@gmail.com; tomas_tat@ukr.net

Application of clusterization for processing of results of the environmental control

The proposed method of processing indicators of sensor sensors of environmental monitoring based on the principle of similarity on the basis of the Ward method.

Екологічний моніторинг є надзвичайно важливим для управління та регулювання навколишнього середовища. З системної точки зору, екологічний моніторинг – це інформаційна система, що займається спостереженням, оцінкою та прогнозом змін у навколишньому середовищі, які виникають при впливі антропогенної складової на природні процеси.

Система екологічного моніторингу накопичує, систематизує та аналізує інформація про: екологічний стан; причини спостережуваних і ймовірних змін в стані навколишнього середовища (джерела та фактори впливу); допустимість змін і навантажень на навколишнє середовище; існуючий потенціал стійкості біосфери.

Першим етапом моніторингу є спостереження навколишнього середовища, отримання та накопичення результатів. В роботах [1,2] був запропонований метод організації розміщення сенсорів на площині довільної форми.

Наступним логічним етапом моніторингу повинна стати обробка результатів спостереження з подальшим аналізом і оцінкою стану навколишнього середовища.

Однак, результатом спостережень зазвичай є набір певних значень параметрів в окремих точках площини. Побудова областей, що складається зі значень однакового рівня в даному випадку є досить складною задачею через наявність випадкової компоненти у виміряних параметрах. Тому задача агрегації (узагальнення) результатів спостережень є актуальною і потребує розробки спеціальних методів.

Для цього пропонується метод кластеризації, який дає можливість об’єднати в агреговані сукупності близькі значення параметрів. Перевагою даного методу є можливість об’єднувати об’єкти, що характеризуються показниками різної фізичної природи (це може бути просторові координати, значення показників сенсорів тощо).

Проведення класифікації об’єктів потребує введення деякої міри подібності (відмінності) об’єктів. Для кількісної характеристики міри подібності використовують поняття відстані між об’єктами в багатовимірному просторі. Подібність або відмінність між об’єктами встановлюється в залежності від відстані між ними.

В нашому випадку як міру відстані між об'єктами застосовуємо евклідову відстань. Якщо об'єкт описується *p* кількісними ознаками, то він може бути представлений як точка *p*-вимірного простору. Тоді евклідова відстань є реальною геометричною відстанню між об'єктами в багатовимірному просторі. Вибір даного вимірника відстані пов'язаний з тим, що в такому дослідженні всі ознаки об'єкта однаково важливі для класифікації.

Для усунення впливу на процедуру класифікації вся вихідна інформація була пронормована відносно середніх показників. Нормування показників *x* проводилось за такою формулою:

$$x_{ij}^{\text{H}} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \qquad (1)$$

де x^H_{ij} - нормоване значення *j*-ої ознаки у *i*-го об'єкта; x_{ij} – значення *j*-ої ознаки у *i*-го об'єкта; $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ – середнє арифметичне значення *j*-ої ознаки; $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ – дисперсія *j*-ої ознаки; *j* та *i* – пробігають значення по стовпцях та рядках, відповідно.

Сутність методу полягає в такому: за допомогою кластерного аналізу набір об'єктів розділяється на основні класи, які утворюють зони з подібними характеристиками навколишнього середовища.

Для проведення кластеризації необхідно визначити правило для об'єднання кластерів, тобто метод, що дає можливість встановити відстань між кластерами. Одним з найефективніших агломеративних методів є метод Уорда [1], який призводить до створення кластерів приблизно рівних за розмірами, що мають форму гіперсфер. Метод Уорда мінімізує внутрішньогрупову суму квадратів (ВСК) відхилень для будь-яких двох кластерів, що можуть бути сформовані на кожному кроці. Тобто розглядається сума квадратів відстаней між кожною точкою (об'єктом, показником) і середньою по кластеру, що містить дану точку. В разі об'єднання кластерів *I* (n_1 елементів) і *J* (n_2 елементів) збільшення внутрішньогрупової суми квадратів змінюється на величину

$$W_{ij} = \frac{n_i \cdot n_j}{n_i + n_j} d_{ij}^2, \qquad (2)$$

де $d_{ij}^2 = (\overline{X^i} - \overline{X^j})^T (\overline{X^i} - \overline{X^j})$;

$\overline{X^i}$, $\overline{X^j}$ – середні значення по кластерах *I* та *J*.

На кожному кроці об'єднуються ті два кластери, для яких збільшення внутрішньогрупової суми квадратів є мінімальним.

Остаточно алгоритм кластеризації набуває такого вигляду:

1. Вважається, що кожний об'єкт утворює свій власний клас. Складається матриця евклідових відстаней $D = \{d_{ij}^2, i = 1,...,p;\ j = 1,...,p\}$ (рис. 1).

D =

	I_1	I_2	…	I_p
I_1	0	d_{12}^2	…	d_{1p}^2
I_2		0	…	d_{2p}^2
…	…	…	…	…
I_p				0

Рис. 1. Матриця евклідових відстаней.

2. Визначається $d_{ML}^2 = min\{d_{ij}^2\},\ i = 1,...,j-1;\ j = 2,...,p$.

3. Збільшення ВСК при об'єднанні двох кластерів I_M та I_L розраховується за формулою

$$W_{ML} = \frac{n_M \cdot n_L}{n_M + n_L} d_{ML}^2, \tag{3}$$

4. I_M змінюється на $I_M^{'}$; рядок $\{d_{iM}^2\}$ та стовпець $\{d_{M\,j}^2\}$ матриці D перераховується за формулою (4)

$$\begin{aligned} d_{iM}^2 = 2W_{iM} &= \frac{2}{n_i + n_M}\left[W_{iM}(n_i + n_M) + W_{iL}(n_i + n_L) - W_{ML}n_i\right] = \\ &= \frac{1}{n_i + n_M}\left[d_{iM}^2(n_i + n_M) + d_{iL}^2(n_i + n_L) - d_{ML}^2 n_i\right], \end{aligned} \tag{4}$$

i=1,2, …, M-1; n_i>0; j=M+1, …, p; $j \neq L$; n_j>0.

5. Припустимо, що $n_M = n_M + n_L$ і $n_L = 0$, тоді кластер I_L перетворюється на недійсну множину.

6. Записуємо елементи кластера I_L в кластер $I_M^{'}$.

7. Повертаємось до пункту 2 і повторюємо процедуру k-2 рази, де k – кількість кластерів, на яку потрібно поділити групу об'єктів.

Оскільки даний метод легко програмується, його застосування дасть можливість швидко агрегувати дані спостережувального етапу моніторингу і на їх основі здійснювати аналіз і прогнозування стану навколишнього середовища.

Література

1. *Гордійко Н.О., Томашевська Т.В.* Використання геометричних фракталів для екологічного моніторингу / Н.О. Гордійко, Т.В. Томашевська. Науковий вісник академії муніципального управління. Збірник наукових праць. Серія "Техніка". Випуск 1-2(11) – К.:, АМУ, 2016. – с.97-104.

2. *Гордійко Н.О., Лисенко О.І., Томашевська Т.В.* Застосування методу використання геометричних фракталів для оптимального розміщення сенсорів при екологічному моніторингу / Н.О. Гордійко, О.І. Лисенко, Т.В. Томашевська. Зб. матеріалів XI Міжнар. наук.-тех. конференції «Проблеми телекомунікацій» ПТ-2017, (м.Київ, 18–21 квітня 2017), с.392-394.

3. *Дюран Б., Одедд П.* Кластерный анализ / Б. Дюран, П. Одедд. М.: Статистика, 1977. 128 с.