УДК 681.3(045)

ОСОБЛИВОСТІ ВИБОРУ ПАРАМЕТРІВ ДИНАМІЧНОЇ КЛАСТЕРИЗАЦІЇ ПРИ АНАЛІЗІ ЕКОЛОГІЧНОГО СТАНУ

Струбицький П.Р.
Тернопільський національний економічний університет, к.т.н., доцент

Динаміка екологічного стану, особливо в місцях з високою ймовірністю можливого забруднення відзначається складністю організації обліку впливу на загальний екологічний стан окремих показників. Тому для ефективного передбачення змін екологічного стану, відстедення динаміки зміни потрібно мати відповідну систему. Ефективне вирішення задач оперативного збору інформації про екологічний стан, прогнозування динаміки його зміни від поєднання взаємного впливу небезпечних факторів неможливе без автоматизації відповідних процесів. Розробці інформаційного забезпечення і математичних моделей аналізу поточного стану та прогнозування розвитку нині приділяється значна увага.

Визначення поточного, а також кінцевого екологічного стану є основним при аналізі екологічних ситуацій. Це може проводитись такими методами:

1. Дані опрацьовує і аналізує експерт, або група експертів.
2. Аналіз даних проводиться на основі інформаційно-довідкових таблиць.
3. Дані аналізуються шляхом аналітичних залежностей між вимірюваними параметрами.

При застосуванні цих методів до розв'язання задачі аналізу динаміки екологічного стану, основними проблемами є низька швидкодія та вплив суб'єктивних факторів. Це визначає необхідність автоматизованого вирішення поставленої задачі.

На сьогоднішній день відсутні єдині підходи до оцінювання екологічної обстановки. Деже поширені методи на основі створення різноманітних математичних моделей, які, проте приймають різну форму та параметри для різних територій, на яких проводиться екологічний моніторинг.

Доцільним при аналізі інформації екологічного моніторингу є використання підходу, що базується на визначенні аналітичних залежностей між параметрами моніторингу. Його перевагою є висока точність, а недоліком – низька швидкодія, що можна усунути шляхом використання новітньої обчислювальної техніки та технологій інтелектуального аналізу даних.

Кластеризація, як одна з основних задач інтелектуального аналізу даних, є способом групування багатовимірних об'єктів, якими є екологічні стани територій. Кластерізація характеризується ітераційним пошуком оптимального рішення, можливістю вибору інформативних ознак та мір, побудовою науково обґрунтованої класифікації багатовимірних спостережень на підставі сукупності відібраних показників та виявлення внутрішніх зв'язків між екологічними станами. Алгоритми кластеризації, на відміну від статистичних методів, можуть бути використані в умовах відсутності інформації про закони розподілу даних екологічного моніторингу.

Нехай Y – множина екологічних станів $Y = \{Y_i\}$ $(i = \overline{1,n})$ (n – потужність множини екологічних станів), що представлена матрицею, в якій кожен рядок $\{y'_{ij}\}$ $(i = \overline{1,n}; j = \overline{1,q})$– множина всіх можливих характеристик, що описує екологічний стан, y'_{ij} – певна характеристика окремого стану (де q – кількість усіх можливих параметрів та характеристик екологічного стану):

$$Y = \{Y_1, Y_2, \dots, Y_n\} = \begin{Bmatrix} y'_{11} & \cdots & y'_{1q} \\ \cdots & \cdots & \cdots \\ y'_{n1} & \cdots & y'_{nq} \end{Bmatrix},$$

де Y_i – i-тий екологічний стан (результат моніторингу на певній території в певний час), y'_{ij} – значення конкретного j-го параметру i-го екологічного стану, q – кількість усіх можливих параметрів екологічного стану, що збережені в базі даних екологічного моніторингу.

З множини усіх параметрів, що аналізується, y' потрібно обрати таку підмножину у, яка б максимально достовірно відображала подібність або відмінність між станами та забезпечувала б достовірне розбиття множини Y станів на k підмножин з дотриманням вимог кластеризації: стан Y_i належить одній і тільки одній підмножині; стани, що належать одній підмножині повинні бути максимально подібними; стани, що належать різним підмножинам повинні бути максимально несхожими, що дозволить проводити аналіз, який ідентифікує стан або динаміку його зміни.

Вибір параметрів y_{ij} при кластеризації є основним з етапів такого аналізу. Основна проблема полягає в тому, що потрібно з множини усіх можливих параметрів екологічного стану вибрати таку їх

підмножину, яка б максимально достовірно відображала подібність або відмінність між станами. Однак, дуже часто виникає проблема "прокляття розмірності" - зі збільшенням множини станів час роботи алгоритму зростає. Однією із причин цього є опрацювання великої кількості надлишкових параметрів, тобто таких, що не несуть корисної інформації, а лише збільшують час роботи алгоритму не впливаючи на його точність.

В ідеальному випадку параметри екологічного стану мають обиратися згідно теорії правильного вибору параметрів, які будуть достовірно відображати міри близькості та відстані між станами [1]. При кластерному аналізі екологічного стану доцільно використати "наївний емпіризм", оскільки даний метод використовується для отримання "об'єктивного" групування множини об'єктів. Хоча емпіричні дослідження важливі при аналізі екологічних станів, але їх використання при кластеризації може викликати появу недостовірних рішень та збільшення часу аналізу даних [2].

Екологічний стан території в конкретний момент часу характеризується повною множиною параметрів, серед них є велика кількість таких, що при кластерному аналізі не впливатимуть на достовірність результатів, а лише збільшуватимуть час роботи алгоритму. Для забезпечення оптимальності роботи алгоритму потрібно вибрати з множини можливих параметрів станів, таку їх підмножину, що мала б мінімальну потужність та максимально точно відображала їх стан. Такою множиною може бути наступна: y_{i1} – концентрація в повітрі двооксиду сірки SO_2; y_{i2} – концентрація в повітрі окису азоту NO; y_{i3} – концентрація в повітрі двооксиду азоту NO_2; y_{i4} – концентрація в повітрі закисів азоту NO_x; y_{i5} – концентрація в повітрі окису вуглецю CO; y_{i6} – концентрація в повітрі озону O_3; y_{i7} – концентрація в повітрі твердих частинок; y_{i8} – швидкість вітру; y_{i9} – напрямок вітру; y_{i10} – температура повітря; y_{i11} – атмосферний тиск; y_{i12} – вологість повітря; y_{i13} – кількість опадів; y_{i14} – сумарна сонячна радіація.

Оскільки, дані екологічного моніторингу мають параметри та характеристики з різними одиницями вимірювання їх потрібно привести до стандартизованого вигляду, особливо при використанні евклідових відстаней. Нормування представляє собою перехід до певного однакового опису всіх параметрів та характеристик екологічних станів, до введення нової умовної одиниці вимірювання, яка допускає формальне співставлення таких станів [1-2].

В [2] зазначено, що змінні багатовимірних даних, якими є параметри екологічного стану, можуть змінювати параметри розподілу від кластера до кластера. Хоча, при кластерному аналізі таких станів нормування може призвести до зменшення відмінностей між кластерами за тими змінними, за якими у вхідному векторі спостерігались значні відмінності. Тому оптимальним є використання декількох способів нормування для різних змінних.

Результати запропонованого підходу до кластеризації екологічного стану було перевірено на даних моніторингу якості повітря в Воєводстві Сілезія республіки Польща. Кількість вимірюваних параметрів достатньо велика і провести ручний аналіз їх в комплексі практично неможливо. Результати вимірювань зберігаються в базі даних WIOS в Катовіце, де вони перевіряються, використовуються і піддаються аналізу. Також дані доступні для всіх дослідників на сайті організації [3]. Результати проведення моніторингу представлені у вигляді таблиць і графіків, як для кожної станції, так і для кожного з вибраних параметрів. Заміри проводяться що години, тому містять достатньо деталізовану інформацію.

За результатами наданих моніторингів за якістю повітря була побудована база даних для дослідження використання кластеризації при аналізі екологічного стану. В результаті були отримані чітко виражені кластери нормального і аномального екологічного стану. Аномальні стани збігалися з результатами аналізів, які проводили працівники лабораторії при оцінці різноманітних загроз та змін в якості повітря. Крім того була отримана динамічна картина зміни кластерів у часі, що досить ефективно показує зміну екологічного стану в реальному часі.

Отже, вибір параметрів при аналізі екологічного стану є одним із важливих етапів кластеризації, оскільки використання даних, які рекомендуються емпірично при цьому може призвести до збільшення часу роботи алгоритму навіть при незначному збільшенні потужності множини таких станів. Кластерний аналіз екологічних станів територій передбачає використання параметрів з різними одиницями вимірювання та потребує приведення їх до нормованого вигляду, що особливо актуально при використанні такої міри близькості як евклідова відстань.

Список використаних джерел

1. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статистика, 1988. – 176с.
2. Дж. – О. Ким. Факторный, дискриминантный и класерный анализ / Дж. – О. Ким, Ч. У. Мюллер, У. Р. Клекка и др.. – М.: Финансы и статистика, 1989. – 216 с.
3. http://stacje.katowice.pios.gov.pl/monitoring/