УДК 519.254

О.П. Приставка, Н.М. Єрещенко, М.Г. Сидорова

Дніпропетровський національний університет імені Олеся Гончара

КЛАСТЕРНИЙ АНАЛІЗ ГІДРОГЕОХІМІЧНИХ ДАНИХ

Запропонована інформаційна технологія проведення кластерного аналізу гідрогеохімічних даних та обчислювальні схеми кластеризації на основі методів: ієрархічної кластеризаціїї, швидкої ієрархічної кластеризації, К-середніх у варіантах Болла-Холла та Мак-Кіна.

Ключові слова: інформаційна технологія, кластерний аналіз, гідрогеохімічний моніторинг, підтримка прийняття рішень.

Предложена информационная технология кластерного анализа данных гидрогеохимического мониторинга и вычислительные схемы кластеризации на основе методов: иерархической кластеризации, быстрой иерархической кластеризации, К-средних в вариантах Болла-Холла и Мак-Кина.

***Ключевые слова*:** информационная технология, кластерный анализ, гидрогеохимический мониторинг, поддержка принятия решений.

Proposed information technology of cluster analysis of hydrogeochemical data monitoring and computational schemes based on clustering methods: hierarchical clustering, fast hierarchical clustering, K-means in the Ball Halls and McKeans variants.

***Keywords**:* information technology, cluster analysis, hydrogeochemical monitoring, support in approval of decisions**.**

Вступ. Навколишнє природнє середовище України стрімко змінюються. Це зумовлює необхідність врахування техногенезу, як головного чинника впливу на довкілля. Проблема изміни та перевищення гранично допустимих концентрацій гідрогеохімічних показників підземної гідросфери є досить актуальною. Тому аналіз гідрогеохімічних даних є важливою складовою для оцінки впливу техногенного навантаження. В якості аназізу розглядаються алгоритми кластерного аналізу, які дають змогу поділити сукупність об'єктів, у даному ипадку свердловин на однорідні за певним формальним критерієм подібності групи (кластери). Основною властивістю цих груп є те, що свердловини, які належать одному кластеру, подібніші між собою, ніж свердловини з різних кластерів.

© А. Ф. Приставка, Н. М. Єрещенко, М. Г. Сидорова, 2010

Постановка задачі. Задано результати гідрогеохімічного дослідження концентрацій хімічних елементів підземних вод у свердловинах на території Північного гірничо-збагачувального комбінату (ПівнГЗКу). Проводилися у свердловинах заміри концентрації хімічних показників:

Cl, SO_4, Ca, Mg, Na, HCO_3, S, SO.

Дані гідрогеохімічного моніторингу представлені у вигляді матриці дійсних чисел

$$X = \{x_{ij}; i = \overline{1,n},\ j = \overline{1,p}\},$$

де

n – кількість свердловин на території ПівнГЗКу, кожна свердловина характеризується набором з p ознак;

x_{ij} – значення вмісту j-ї хімічної речовини в i-й свердловині.

Необхідно розробити інформаційну технологію проведення кластерного аналізу гідрогеохімічних даних та запропонувати обчислювальні схеми кластеризації на основі методів: ієрархічної кластеризаціїї, швидкої ієрархічної кластеризації, К-середніх.

Розбиття на кластери, що отримані різними методами, або при різних значеннях параметрів, можуть відрізнятися. Тому слід провести оцінку якості отриманих результатів на основі функціоналів якості. Для визначення методу, що дає найкраще розбиття на кластери необхідно реалізувати методи підтримки прийняття рішень, а саме множинний аналіз, процедури Борда та плюралітарну. Де експертами будуть виступанти функціонали якості, а альтернативами – методи кластеризації.

Основний матеріал. Для розв’язання поставленої задачі запропоновано інформаційну технологію, структурна схема якої представлена на рис.1.

Якщо ознаки об’єктів вимірюються у різних одиницях або сильно відрізняються за значеннями, доцільно привести їх до єдиного масштабу, тобто провести стандартизацію.

Розглянемо основні етапи інформаційної технології проведення кластерного аналізу гідрогеохімічних даних.

Під час проведення кластерного аналізу важливим і найменш формалізованим є поняття близькості об’єктів, що визначається деякою функцією $d(i, j) \geq 0$ [2]:

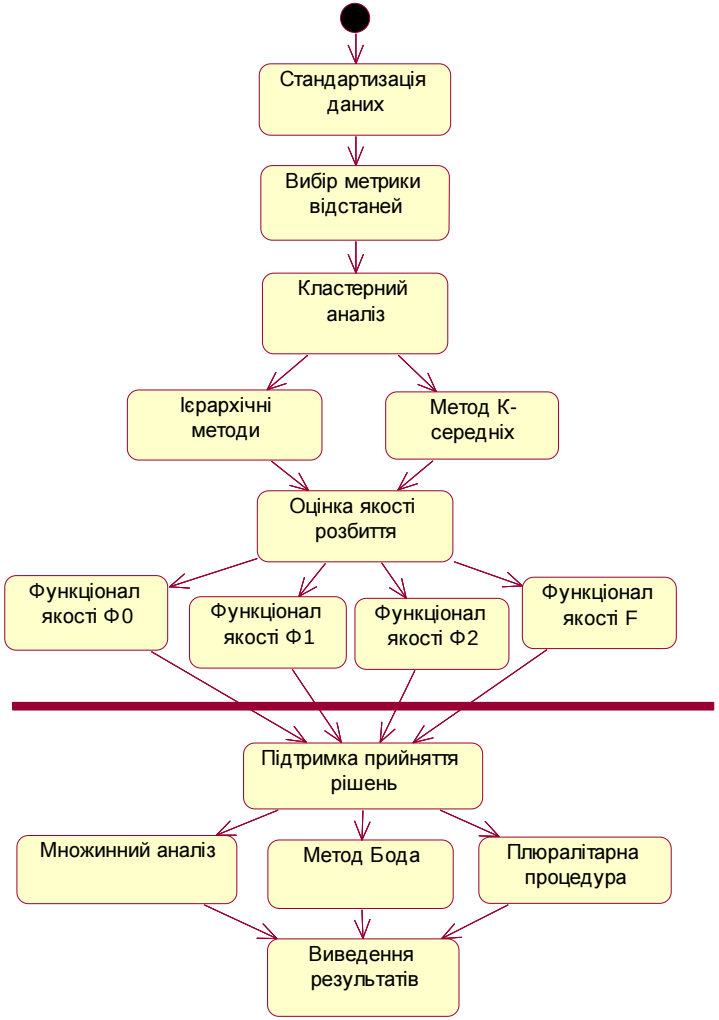

Рис. 1. Загальна схема інформаційної технології гідрогеохімічного моніторингу

$$d(i,j) \le d(i,k) + d(k,j), \quad \forall i,j,k;$$
$$d(i,j) = d(j,i), \quad \forall i,j;$$
$$d(i,j) = 0, \quad i = j.$$

В залежності від того, яким способом обчислюється відстань між об'єктами значною мірою залежить результат, тому існує безліч варіанти метрик, найбільш поширені з яких:

- евклідова відстань відповідає простій геометричній відстані в багатовимірному просторі

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2};$$

- манхетенська відстань (відстань міських кварталів) ефективно використовується для бінарних ознак, але також може бути застосована і для якісних ознак

$$d(i,j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|;$$

- відстань Чебишева ефективно використовується у тому випадку, коли існує одна ознака, за якою можливо здійснити ефективне розділення класів

$$d(i,j) = \max_{1 \le k \le p} \left| x_{ik} - x_{jk} \right|.$$

Також важливою величиною в кластерному аналізі є відстань між цілими групами об'єктів [1]. Розглядаються найбільш поширені відстані і міри близькості, що характеризують взаємне розташування окремих груп об'єктів. Нехай відстань $d(C_l, C_m)$ між двома кластерами C_l і C_m з кількістю елементів n_l та n_m відповідно можливо визначити як:

- відстань найближчого сусіда

$$d(C_l, C_m) = \min_{\substack{X_i \in C_l; \\ X_j \in C_m}} d(X_i, X_j);$$

- відстань найвіддаленішого сусіда

$$d(C_l, C_m) = \max_{\substack{X_i \in C_l; \\ X_j \in C_m}} d(X_i, X_j);$$

- середня відстань

$$d(C_l, C_m) = \frac{1}{n_l n_m} \sum_{X_i \in C_l} \sum_{X_j \in C_m} d(X_i, X_j);$$

- відстань між центрами $d(C_l, C_m) = d(M_l, M_m)$,
де
M_i – центр і-го класу;
- відстань Уорда

$$d(C_l, C_m) = \frac{n_l n_m}{n_l + n_m} d^2(M_l, M_m).$$

Проведедення кластерного аналізу запропоновано наступними методами:

- ієрархічним агломеративним методом [1 – 4], швидким ієрархічним метод [2] та методом К-середніх [1 – 2].

Ієрархічний агломеративний метод характеризується послідовним об'єднанням вихідних елементів та відповідним зменшенням числа кластерів. На початку роботи алгоритму всі об'єкти є окремими кластерами. На першому кроці найбільш схожі об'єкти об'єднуються в кластер. На наступні кроки об'єднання продовжується до тих пір, поки всі об'єкти не будуть складати один кластер.

Пропонується обчислювальна схема ієрархічного агломеративний методу:

Кожен обєкт вважається окремим кластером, для отриманих одноелементних кластерів обчислюється метрика відстані

$$D = \{d_{ij}\}, \quad i, j = \overline{1,n},$$

де

$$d_{ij} = d(i, j)$$

.

На кожній ітерації обєднуються в один кластер кластери, для яких $d_{ij} \to \min$. Після чого, з матриці D вилучають відповідні кластери. Процедуру слід повторювати $n-1$ раз, доки всі об’єкти не будуть об’єднані.

Для обчислення відстані між кластерами існує загальна формула Ланса-Уільямса:

$$d(C_{l+m}, C_h) = \alpha_l \cdot d(C_l, C_h) + \alpha_m \cdot d(C_m, C_h) + \beta \cdot d(C_l, C_m) + \\ + \gamma \cdot |d(C_l, C_h) - d(C_m, C_h)|$$

Задаючи різні значення параметрів α_l, α_m, β, γ, отримаємо різні види агломеративних ієрархічних методів:

- $\alpha_l = \frac{1}{2}; \quad \alpha_m = \frac{1}{2}; \quad \beta = 0; \quad \gamma = -\frac{1}{2}$ – найближчого сусіда;
- $\alpha_l = \frac{1}{2}; \quad \alpha_m = \frac{1}{2}; \quad \beta = 0; \quad \gamma = \frac{1}{2}$ – найвіддаленішого сусіда;

- $\alpha_l = \frac{n_l}{n_l + n_m}; \quad \alpha_m = \frac{n_m}{n_l + n_m}; \quad \beta = 0; \quad \gamma = 0$ – середнього зв’язку;
- $\alpha_l = \frac{n_l}{n_l + n_m}; \quad \alpha_m = \frac{n_m}{n_l + n_m}; \quad \beta = -\frac{n_l n_m}{(n_l + n_m)^2}; \quad \gamma = 0$ – між центрами;
- $\alpha_l = \frac{n_h + n_l}{n_h + n_l + n_m}; \quad \alpha_h = \frac{n_h + n_m}{n_h + n_l + n_m}; \quad \beta = -\frac{n_h}{n_h + n_l + n_m}; \quad \gamma = 0$ – Уорда.

Ідея швидкого ієрархічного методу кластеризації полягає в тому, щоб перебирати лише найбільш близькі пари. Тобто для економії пам’яті і зменшення необхідного числа порівнянь при пошуку найменшої відстані слід виключити з розгляду ті відстані, які не впливають на виконання обчислень.

Обчислювальна схема методу:

Нехай обрано метрику відстані й обчислено матрицю відстаней $D = \{d_{ij}\}, i, j = \overline{1,n}$, де $d_{ij} = d(i, j)$. Кожен об’єкт вважається окремим кластером.

Обирається параметр δ. В залежності від способу обчислення δ існують різні методи швидкої кластеризації.

Метод1. В якості δ обирають середнє значення матриці відстаней.

Метод2. Задаються параметри n1 та n2. Якщо число кластерів не перевищує поріг n1, то δ не обирають, а працюють з матрицею D. Інакше вибирається випадковим чином n2 відстаней, і δ обирається рівним найменшому з них. Параметри n_1 та n_2 впливають лише на час виконання алгоритму, а не на результат кластеризації. В якості початкового вибору можна запропонувати $n_1 = n_2 = 20$.

Формується множина пар $P(\delta) : \{(C_l, C_m) : d(C_l, C_m) \le \delta\}$. Далі алгоритм співпадає з алгоритмом класичної агломеративної кластеризації, тільки замість матриці D використовується множина $P(\delta)$.

Коли всі такі пари будуть вичерпані, параметр δ збільшується, і формулюється нова скорочена множина пар. Так продовжується до повного злиття всіх об’єктів в один кластер.

Результати швидких методів ієрархічної кластеризації повністю співпадають з класичними ієрархічними методами. Відмінність полягає лише у швидкості роботи.

Метод К-середніх – алгоритм розділової кластеризації, заснований на розбитті елементів векторного простору на певне число кластерів k. Полягає у послідовному уточненні точок центру кластерів.

Алгоритм представляє собою ітераційну процедуру, в якій виконуються наступні кроки:

На нульовому наближенні задається кількість кластерів та серед множини точок обирається K точок, які вважаються центром кластеру, одним із запропонованих способів:

- перші K точок;
- випадкові K точок;
- найвіддаленіші K точок.

Для кожної точки $X_i, i=\overline{1,n}$ визначається найближчий до неї центр кластеру. При цьому, точки, «притягнуті» певним центром, утворюють початкові кластери.

Обчислюються відстані $d(X_i, M_j), j=\overline{1,K}$ та обирається номер того кластеру, де буде досягатися мінімум.

Обчислюється центроїда – центри тяжкості кластерів. Кожен центроид – це вектор $M_i=(\bar{x}_{i1}, \bar{x}_{i2},..., \bar{x}_{ip}), i=\overline{1,K}$, елементи якого являють собою середні значення ознак $\bar{x}_{ij}=\frac{1}{n_i}\sum_{k=1}^{n_i} x_{kj}, j=\overline{1,p}$, обчислені за всіма записами кластеру. Потім центр кластера зміщується в його центроид.

Третій та четвертий кроки ітеративно повторюються. Очевидно, що на кожній ітерації відбувається зміна меж кластерів і зсув їхніх центрів. У результаті мінімізується відстань між елементами всередині кластерів. Зупинка алгоритму відбувається тоді, коли кордони кластерів і розташування центроїд не перестануть змінюватися від ітерації до ітерації, тобто на кожній ітерації в кожному кластері буде залишатися один і той самий набір записів.

Оцінюється якість розбиття отриманого кожним з розглянутих методів на основі функціоналів якості. Задається кількість кластерів = 4, а метрика відстані – евклідова. Нехай у результаті застосування методу кластеризації одержане розбиття $C=\{C_1, C_2,..., C_K\}$ множини об’єктів на K кластерів.

$$C_i=\{X_1^{(i)}, X_2^{(i)},..., X_{n_i}^{(i)}\}=\begin{pmatrix} x_{11}^{(i)} & x_{12}^{(i)} & ... & x_{1i}^{(i)} & ... & x_{1j}^{(i)} & ... & x_{1p}^{(i)} \\ x_{21}^{(i)} & x_{22}^{(i)} & ... & x_{2i}^{(i)} & ... & x_{2j}^{(i)} & ... & x_{2p}^{(i)} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ x_{n_l1}^{(i)} & x_{n_l2}^{(i)} & ... & x_{n_li}^{(i)} & ... & x_{n_lj}^{(i)} & ... & x_{n_lp}^{(i)} \end{pmatrix}$$

– і-й кластер.

Існує багато різновидностей функціоналів якості кластеризації [1, 2, 4], розглянемо найбільш розповсюджені:

- сума внутрішньокластерних відстаней

$$\Phi_0=\sum_{l=1}^{K}\sum_{i=1}^{n_l-1}\sum_{j=i+1}^{n_l} d(X_i^{(l)}, X_j^{(l)}) \to \min;$$

- сума квадратів відстаней до центрів кластерів

$$\Phi_1=\sum_{i=1}^{K}\sum_{X_j\in C_i} d^2(X_j^{(i)}, M_i) \to \min,$$

де

$M_i=(\bar{x}_1^{(i)}, \bar{x}_2^{(i)},..., \bar{x}_p^{(i)})$ – центр кластера C_i;

- сума внутрішньокластерних дисперсій за всіма ознаками

$$\Phi_2=\sum_{i=1}^{K}\sum_{j=1}^{p}\left(\frac{1}{n_i-1}\sum_{h=1}^{n_i}\left(x_{hj}^{(i)}-\bar{x}_j^{(i)}\right)^2\right) \to \min;$$

- відношення функціоналів

$$F=\frac{F_0}{F_1} \to \min,$$

де

середня внутрішньокластерна відстань

$$F_0=\frac{1}{\sum_{i=1}^{K}\frac{n_i(n_i-1)}{2}}\sum_{i=1}^{K}\sum_{j=1}^{n_i-1}\sum_{h=j+1}^{n_i} d(X_j^{(i)}, X_h^{(i)}),$$

середня міжкластерна відстань

$$F_1=\frac{1}{\prod_{i=1}^{K} n_i}\sum_{i=1}^{K-1}\sum_{j=1}^{n_i}\left(\sum_{m=i+1}^{K}\sum_{h=1}^{n_m} d(X_j^{(i)}, X_h^{(m)})\right).$$

Результати обчислень, попередньо зведені до одиничної шкали

представлені у таблиці 1.

Таблиця 1

Оцінки якості методів кластерного аналізу за функціоналами якості

Назва методу	Φ_0	Φ_1	F	Φ_2
Ієрархічні:				
ближнього сусіда	1,00	1,000	0,045	1,000
дальнього сусіда	0,748	0,655	0,079	0,68
середня відстань	0,748	0,655	0,079	0,68
відстань Уорда	0,482	0,551	1,000	0,207
Неієрархічні:				
К-середніх:				
перші К точок	0,539	0,489	0,787	0,199
найвіддаленіші К точок	0,689	0,489	0,293	0,298
випадкові К точок	0,539	0,489	0,787	0,198

Визначити найкраще розбиття за допомогою функціоналів якості досить складно. Оскільки найчастіше функціонали, через те, що до кожного з них закладено різні поняття кластера та однорідності, обирають різні методи кластеризації як найкращі. Таким чином виникає неоднозначність результатів та потреба у реалізації процедур прийняття рішень. В ролі експертів будемо розглядати функціонали якості, а альтернативами будуть методи кластерного аналізу. Дані представимо у вигляді матриці $X = \{x_{ij}; i = \overline{1,n}, j = \overline{1,m}\}$, де n – кількість методів, m – кількість експертів, x_{ij} – оцінка, яку поставив j-й експерт і-й альтернативі. Розглянемо 3 методи прийняття рішень: множинний аналіз, процедура Борда та плюралітарна процедура [5].

Обчислювальна схема множинного аналізу:

1) Задаємо $t = 1, k_j = \frac{1}{m}$.

2) Обчислюємо $x_i^t = \sum_{j=1}^{m} x_{ij} k_j^{t-1}, i = \overline{1,n}$.

3) Обчислюємо $\lambda^t = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} x_i^t, t = 1,2,...$

4) Збільшуємо $t : t = t + 1$. Обчислюємо значення $k_j^t = \frac{1}{\lambda^t} \sum_{i=1}^{n} x_{ij} x_i^t, \sum_{j=1}^{m} k_j^t = 1, j = \overline{1,m}$.

5) Повторюємо пункти 2 – 4, до тих пір поки процес не зійдеться з заданою точністю ε.

Оцінку якості методів кластеризації за множинним аналізом представлено у таблиці 2.

Таблиця 2

Оцінки якості методів кластерного аналізу

Методи	*Оцінка*
Ієрархічні:	
ближнього сусіда	0,8267
дальнього сусіда	0,5854
середня відстань	0,5854
відстань Уорда	0,5372
К-середніх:	
перші К точок	0,4972
випадкові К точок	0,4972
найвіддаленіші К точок	0,4631

Процедура Борда. Оцінки кожного експерта впорядковуються. Обчислюється оцінка для кожної альтернативи, як сума рангових місць. Найкращим варіантом вважається той, що буде мати найменшу оцінку.

Плюралітарна процедура. Оцінки кожного експерта впорядковуються. Кожній альтернативі присвоюється оцінка, що дорівнює кількості експертів, які поставили її на перше місце. Найкращою вважається альтернатива з максимальною оцінкою.

Висновки. Розроблено інформаційну технологію кластерного аналізу даних гідрогеохімічного дослідження стану води у свердловинах на території гірничо-збагачувального комбінату. Запропоновано обчислювальні технології методів кластеризації, оцінки якості отриманого розбиття на основі функціоналів якості, вибір найкращого методу за допомогою множинного аналізу та

процедур колективного вибору. Створено систему, ядро якої складають запропоновані обчислювальні технології.

Бібліографічні посилання

1. **Айвазян С.А.** Классификация многомерных наблюдений / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М., 1974. – 240 с.
2. **Жамбю М.** Иерархический кластер-анализ и соответствия / М. Жамбю. – М., 1988. – 279 с.
3. **Загоруйко Н.Г.** Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск, 1999. – 270 с.
4. **Мандель И.Д.** Кластерный анализ / И.Д. Мандель. – М., 1988. – 176 с.
5. **Емельяненко Т.Г.** Принятие решений в системах мониторинга / Т.Г. Емельяненко, А.В. Зберовский, А.Ф. Приставка, Б.Е. Собко. – Д., 2005. – 224 с.

Надійшла до редколегії 20.12.2010

Відомості про авторів

Аксьоненко Павло Юрийович – студент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології розпізнавання мовних сигналів.

Байбуз Олег Григорович **–** д-р техн. наук, професор, завідувач кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Бовкун Олександр Сергійович **–** студент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: теорія графів, методи комбінаторної оптимізації, інформаційні технології обробки статистичних даних.

Голуб Сергій Васильович **–** д-р техн. наук, доцент, директор центру Черкаського національного університету ім. Богдана Хмельницького.

Коло наукових інтересів: Системи багаторівневого перетворення моніторингової інформації.

Девяткін Іван Вікторович – аспірант кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Дроздова Ірина Володимирівна **–** д-р медичних наук, провідний науковий співробітник лабораторії експертно-реабілітаційної ультразвукової та функціональної діагностики Українського Державного науково-дослідницького інституту медико-соціальних проблем інвалідності, м. Дніпропетровськ.

Коло наукових інтересів: кардіологія.

Ємел`яненко Тетяна Георгіївна – канд. техн. наук, доцент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.