УДК 519.254

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСТЕРИЗАЦІЇ ДАНИХ У ЧАСОВОМУ ПЕРІОДІ СПОСТЕРЕЖЕНЬ

О.Г. БАЙБУЗ, М.Г. СИДОРОВА

Кластерний аналіз є актуальним напрямом інтелектуального аналізу даних (Data Mining). Застосування методів кластеризації дозволяє зрозуміти структуру багатовимірних даних; спростити подальшу обробку, використовуючи різні методи аналізу для кожного кластера; скоротити вихідну вибірку даних, залишивши по одному найбільш типовому представнику кожної групи; виявити новизну, нетипові об'єкти, які не вдається приєднати до жодного з класів; сформулювати або перевірити гіпотези на підставі отриманих результатів. Запропоновано новий підхід до виділення груп об'єктів, схожих між собою за набором ознак, які змінюються у часі. Розроблено інформаційну технологію оцінки якості й підвищення стійкості кластеризації. Представлено результати практичної реалізації запропонованої технології на даних гідрохімічного моніторингу водних об'єктів у районі з підвищеним техногенним навантаженням.

ВСТУП

На сьогодні все більшої популярності набуває такий напрям обробки інформації як інтелектуальний аналіз даних (Data Mining), до задач якого відноситься також кластерний аналіз, що застосовується для виявлення груп схожих між собою об'єктів, ієрархічних структур і закономірностей у наборі даних. Результатом кластерного аналізу є розбиття (угрупування) об'єктів вихідної вибірки на групи (кластери) таким чином, що об'єкти однієї групи є більш схожими за набором досліджуваних ознак, ніж об'єкти з різних груп. Методи кластеризації широко використовуються в інформаційних технологіях під час роботи з базами даних, аналізі інтернет-документів, сегментації зображень, дослідженнях у медицині, економіці, екології, соціології, психології тощо.

Існує багато різних підходів та методів кластерного аналізу [1–5]. Проте незважаючи на значну кількість досліджень, у цій галузі є ряд актуальних проблем та питань, які не знайшли свого повного розв'язку.

Одним із найактуальніших питань кластерного аналізу є оцінювання результатів та пошук розбиття, що найкраще відповідає структурі даних [5–7]. Як відомо, результат кластеризації досить сильно залежить від вибору системи ознак, мір близькості, способів формалізації уявлень про схожість об'єктів та кластерів. Кластеризаційні схеми, отримані різними методами або за різних значень параметрів можуть значно відрізнятися або не відповідати об'єктивно існуючим угрупуванням. На сьогодні в літературі існує велика кількість функціоналів та індексів якості, що дозволяють у кількісному вигляді оцінювати відповідність отриманого розбиття природній структурі даних, а також порівнювати результати, отримані різними методами або за різних значеннях параметрів. Визначення функціоналів якості головним чином ґрунтується на таких критеріях, як компактність та відокремленість кластерів, але все ж таки до кожного з них закладено різні поняття кластера та однорідності, тому вони

досить часто демонструють зовсім різні результати, «обираючи» різні розбиття як найякісніші. У цій роботі пропонується технологія оцінки якості, яка на основі методів прийняття рішень дозволяє враховувати значення різних функціоналів якості одночасно, що забезпечує більш точну оцінку результатів.

Досить часто виникає задача виділення груп схожих об'єктів за набором ознак, які змінюються у часі, тобто значенню кожної ознаки для кожного об'єкта відповідає не окреме число, а часовий ряд. Тому актуальним напрямом є розробка нових підходів та алгоритмів для розв'язання такої задачі. У цій роботі рекомендується технологія, яка ґрунтується на колективних методах кластеризації та дозволяє виділяти групи схожих об'єктів за набором показників у часовому періоді спостережень.

ПОСТАНОВКА ЗАДАЧІ

Нехай маємо N об'єктів спостереження, які характеризуються p ознаками, значення яких змінюються протягом T моментів часу. Тобто вихідні дані представлено у вигляді $X = \{x_{ijt}\}$, $i = \overline{1, N}$, $j = \overline{1, p}$, $t = \overline{1, T}$, де x_{ijt} — значення j-го показника i-го об'єкта у момент часу t. Необхідно розподілити об'єкти вихідної вибірки на групи (кластери) за схожістю досліджуваних ознак із урахуванням їх часових змін. Тобто отримати угрупування $G = \{g_1, g_2, \ldots, g_K\}$, де K — кількість кластерів, g_i, $i = \overline{1, K}$: $g_i = \{x_l\}$, $l = \overline{1, N_i}$, — i-й кластер, що містить N_i об'єктів вихідної вибірки, $x_l = \{x_{ljt}\}$, $j = \overline{1, p}$, $t = \overline{1, T}$, $\sum_{i=1}^{K} N_i = N$, $\bigcup_{i=1}^{K} g_i = X$, $g_i \cap g_j = \varnothing$, $i, j = \overline{1, K}$, $i \neq j$. Методи кластерного аналізу в якості вихідної інформації використовують матрицю «об'єкти-ознаки» $X = \{x_{ij}\}$, $i = \overline{1, N}$, $j = \overline{1, p}$, тому не можуть бути застосовані, коли значенню кожної ознаки для кожного об'єкта відповідає не окреме число, а часовий ряд.

Мета роботи — розробка нового методу та інформаційної технології, що дозволить виділяти групи об'єктів, схожих між собою за усіма досліджуваними ознаками, які змінюються у часі та враховувати часові зміни досліджуваних показників.

Крім того, технологія має забезпечувати контроль та підвищення якості кластеризації, забезпечувати підтримку прийняття рішень, візуалізацію та інтерпретацію отриманих результатів.

ОСНОВНІ РЕЗУЛЬТАТИ

Як відомо, різні методи кластерного аналізу, застосовані до одного і того ж набору даних, можуть демонструвати досить відмінні результати. Випадковий необґрунтований вибір методу може призвести до того, що отримане ним розбиття буде зовсім відмінним від природної, притаманної досліджуваним даним кластерної структури. Існують різноманітні функціонали та індекси якості, які дозволяють порівнювати отримані різними методами розбиття та обирати найякісніше з них за певним обраним критерієм. Найчастіше використовують функціонали такі, як сума внутрішньокластерних дисперсій

за всіма ознаками, відношення середньої внутрішньокластерної та середньої міжкластерної відстаней, сума квадратів відстаней до центрів кластерів, а також індекси Данна, Беджека, Девіса-Болдуіна тощо.

У цій роботі пропонується технологія багатокритеріальної оцінки якості результатів кластеризації, яка складається з таких етапів:

- Проводимо кластеризацію об'єктів вихідної вибірки різними методами. Таким чином отримуємо набір угрупувань, серед яких необхідно обрати найякісніше.
- Обираємо критерії (функціонали та індекси) якості, за якими будемо оцінювати та порівнювати отримані розбиття.
- Обчислюємо оцінки кожного критерію для кожного варіанту кластеризації. Отримані результати представляємо у вигляді матриці $Q = \{q_{ij};\ i = \overline{1,n},\ j = \overline{1,m}\}$, де n — кількість порівнюваних угрупувань, m — кількість критеріїв оцінки якості, q_{ij} — оцінка якості *i*-го варіанта розбиття за j-м критерієм.
- Задачу визначення найякіснішого за декількома критеріями варіанту розбиття вихідної вибірки об'єктів на кластери можна сформулювати у термінах теорії прийняття рішень. В якості альтернатив будемо розглядати результати кластеризації отримані різними методами, а в якості експертних оцінок — значення функціоналів якості, обчислені для кожної альтернативи, тобто матриця $Q = \{q_{ij};\ i = \overline{1,n},\ j = \overline{1,m}\}$. Методи колективного вибору дозволяють перейти від індивідуальних (за одним критерієм) до узагальнюючих (за всіма критеріями) оцінок порівнюваних альтернатив. Таким чином, ми можемо ранжувати альтернативи, що дає змогу обирати найякісніші розв'язки. Пропонується застосовувати наступні колективні методи прийняття рішень [8]:

– Процедура Борда. Для кожного критерію виконуємо впорядкування альтернатив у порядку спадання їх якості. Обчислюємо колективну оцінку якості альтернативи як суму рангових місць за кожним критерієм. Найкращим результатом вважається той, що буде мати найменшу оцінку.

– Плюралітарна процедура. Ранжуємо альтернативи окремо за кожним критерієм. Для кожної альтернативи обчислюємо колективну оцінку, що дорівнює кількості критеріїв, за якими вона є найякіснішою. Найкращою вважається альтернатива з максимальною оцінкою.

– Множинний аналіз. Перетворюємо виставлені оцінки за наступною формулою: $q_{ij} = \dfrac{q_{ij}}{\sum\limits_{i=1}^{n} q_{ij}}$. Оцінка якості альтернатив проводиться за рекурентною процедурою. На кожному кроці i обчислюємо оцінки альтернатив $r_l^i = \sum\limits_{j=1}^{m} q_{ij} k_j^{i-1},\quad l = \overline{1,n},$ де $k_j^0 = \dfrac{1}{m},\ k_j^i = \dfrac{1}{\lambda^i} \sum\limits_{l=1}^{n} q_{ij} r_l^i,\quad j = \overline{1,m},\quad \sum\limits_{l=1}^{n} k_l^i = 1,$ $\sum\limits_{l=1}^{n} \sum\limits_{j=1}^{m} q_{ij} r_l^i$ доки процес не зійдеться з деякою заданою точністю ε. Доведе-

но, що процес є збіжним [9]. Найкращим вважається результат із мінімальною оцінкою. Цей метод дозволяє також оцінити узгодженість критеріїв на основі дисперсійного коефіцієнта конкордації.

- Іноді декілька результатів кластеризації можуть представляти різні угрупування, але бути рівнозначними за якістю. У такому випадку можна замість вибору одного з цих розв'язків побудувати на їх основі ансамбль алгоритмів та отримати результуючий розв'язок. Ансамблям алгоритмів присвячено багато робіт, зокрема [4, 5, 10, 11].

Для виділення груп об'єктів, схожих між собою за набором ознак, які змінюються у часі, пропонується технологія *часової кластеризації*, що складається з трьох основних етапів: визначення груп об'єктів для кожного моменту часу $t = \overline{1,T}$, формування узагальненої матриці подібності, отримання підсумкового розв'язку задачі.

На першому етапі знаходимо розбиття об'єктів вихідної вибірки на кластери за даними, що визначають кожен з моментів часу $t = \overline{1,T}$. Взагалі, отримати розбиття для певного моменту часу можна будь-яким методом кластерного аналізу. Однак, оскільки результати досить сильно залежать від вибору методу, авторами пропонується застосовувати вищеописану технологію багатокритеріальної оцінки якості кластеризації для вибору найякіснішого розбиття, тобто розбиття, що найкраще відповідає природній структурі досліджуваних даних. Таким чином отримуємо T угрупувань, кожне з яких є результатом кластеризації, що характеризує певний момент часу. На другому етапі переходимо від визначення схожості об'єктів у деякий окремий момент часу до визначення їх подібності у часовому діапазоні. Для цього на основі отриманих угрупувань формуємо узагальнену матрицю подібності. Використовуючи цю матрицю, на завершальному етапі отримуємо результуюче підсумкове розбиття об'єктів на групи, яке враховує часові зміни досліджуваних ознак.

Визначення угрупувань для кожного моменту часу. Представимо вихідні дані у вигляді групи матриць $X^{(t)} = \{x_{ij}^{(t)}\}$, $i = \overline{1,N}$, $j = \overline{1,p}$, $t = \overline{1,T}$. Застосовуючи відомі методи кластерного аналізу [1–5] та технологію багатокритеріальної оцінки якості, визначимо розбиття об'єктів на кластери окремо для кожного моменту часу. Тобто отримаємо T угрупувань $G_t = \{g_1^{(t)}, g_2^{(t)}, \ldots, g_K^{(t)}\}$, $t = \overline{1,T}$, де K — кількість кластерів, $g_i^{(t)}$, $i = \overline{1,K}$ — i-й кластер у t-му угрупуванні, $g_i^{(t)} = \{x_l^{(t)}\}$, $l = \overline{1,N_i^{(t)}}$, $N_i^{(t)}$ — кількість об'єктів у i-му кластері t-го угрупування, $x_i^{(t)} = \{x_l^{(t)}\}$, $j = \overline{1,p}$.

Формування узагальненої матриці подібності. У кластерному аналізі важливим і найменш формалізованим є вибір способу визначення схожості між об'єктами. У загальному випадку ступінь схожості будь-якої пари об'єктів вихідної множини задається або обчисленням відстані між ними на основі деякої метрики, або введенням правила визначення міри близькості.

Таким чином, для часової кластеризації слід визначити міру близькості, що буде характеризувати схожість двох об'єктів у часовому діапазоні. Пропонується в якості міри близькості двох об'єктів вважати нормовану частоту

їх віднесення до одного кластеру протягом T моментів спостереження, тобто $\mu(i,j)=\dfrac{\sum_{t=1}^{T} y_t}{T}$, де $y_t=1$, якщо об'єкти *i* та *j* відносяться до одного кластеру у *t*-му угрупуванні, $y_t=0$ у іншому випадку. Оскільки віднесення об'єктів до одного кластеру у певний момент часу свідчить про їх близькість за досліджуваними ознаками, а частота їх об'єднання вказує на схожість у часі, то таким чином введена міра близькості дійсно відображає ступінь схожості двох об'єктів за набором ознак із урахуванням часових змін.

На основі отриманої множини розв'язків G_t, $t=\overline{1,T}$ та введеного поняття міри близькості формуємо узагальнену матрицю подібності об'єктів $S=\{s_{ij}\}$, $i,j=\overline{1,N}$, де N — кількість об'єктів, s_{ij} — міра близькості *i*-го та *j*-го об'єктів.

Алгоритм формування матриці подібності:

- Створюємо матрицю $S=\{s_{ij}\}$, $i,j=\overline{1,N}$, та ініціалізуємо її нулями: $s_{ij}=0$, $i,j=\overline{1,N}$.
- Розглядаємо по черзі отримані результати кластеризації G_t: $t=\overline{1,T}$ для кожного моменту часу. Якщо *i*-й та *j*-й об'єкти у *t*-му угрупуванні відносяться до одного кластеру, то s_{ij} збільшуємо на одиницю: $s_{ij}=s_{ij}+1$, інакше — значення s_{ij} залишаємо без змін.
- Зводимо елементи матриці подібності до одиничної шкали: $s_{ij}=\dfrac{s_{ij}}{T}$, $i,j=\overline{1,N}$. Після такого перетворення s_{ij} набувають значень на відрізку від 0 до 1. Чим ближче значення s_{ij} до одиниці, тим більш схожими є об'єкти *i* та *j* на всьому часовому проміжку спостереження.

Отримання підсумкового розв'язку. На завершальному етапі необхідно отримати підсумковий розв'язок поставленої задачі, а саме: розбиття об'єктів вихідної множини на кластери. Об'єднаними у відповідні кластери мають бути ті об'єкти, що є схожими між собою за всіма досліджуваними ознаками з урахуванням їх часових змін. Оскільки на попередньому кроці було визначено міру близькості та сформовану матрицю подібності об'єктів у часовому діапазоні, то отримати підсумкове розбиття можна застосовуючи алгоритми кластерного аналізу, які в якості вихідної інформації використовують матрицю відстаней між об'єктами (наприклад, ієрархічні або графові методи). Пропонується застосовувати графовий алгоритм найкоротшого незамкненого шляху, оскільки він є досить простим у реалізації та демонструє хороші результати. Перехід від матриці близькості $S=\{s_{ij}\}$, $i,j=\overline{1,N}$, до матриці відстаней $S'=\{s'_{ij}\}$, $i,j=\overline{1,N}$ можна здійснити таким чином: $s'_{ij}=1-s_{ij}$, $i,j=\overline{1,N}$. Тобто чим більше подібні об'єкти *i* та *j* за матрицею S, тим менша відстань між ними у матриці S'.

ПРАКТИЧНА РЕАЛІЗАЦІЯ

Запропоновану технологію було застосовано до даних гідрохімічного моніторингу, що проводиться Криворізькою геологогідрогеологічною партією по р. Інгулець (Кривбас). Метою роботи було визначення груп пунктів спостереження, що характеризуються схожим хімічним складом води у р.Інгулець за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки.

Об'єктом дослідження є хімічний склад води у р. Інгулець поблизу ВАТ «Центрального гірничо-збагачувального комбінату». Проби води відбиралися у 5 пунктах спостереження: село Тернівка, створи балок — Мала Лозоватка, Велика Лозоватка, Завертана, північна частина Карачунівського водосховища. Аналіз проводився за вмістом головних іонів у воді річки Інгулець: HCO_3^-, $C1^-$, SO_4^{2}, Ca^{2+}, Mg^{2+}, Na^+ та мінералізацією протягом наступних років: 1993–1995, 1997, 2001, 2003, 2005–2007.

Таким чином маємо п'ять об'єктів (пункти спостереження), кожен з яких характеризується сімома ознаками (вміст іонів у воді річки), значення яких вимірюються дев'ять разів. Тобто вихідні дані можна представити у вигляді $X = \{x_{ijt}\}$, $i = \overline{1,5}$, $j = \overline{1,7}$, $t = \overline{1,9}$, де x_{ijt} — значення j-го показника i-го об'єкта у момент часу t. Для зведення даних до єдиного масштабу попередньо проведена стандартизація.

Сформуємо та розглянемо по черзі дев'ять матриць «пункти спостереження — значення досліджуваних ознак», кожна з яких відповідає певній даті відбору проб води з річки. За допомогою методів кластерного аналізу (ієрархічних: одиничного, повного, середнього зв'язку, Уорда; К-середніх: Болла-Холла, Мак-Кіна; графового, Forel) та запропонованої технології багатокритеріальної оцінки якості результатів отримуємо угрупування схожих між собою об'єктів для кожного окремого моменту часу.

Розглянемо, наприклад, розбиття, що відповідають даним 2001 та 2007 років (рис. 1, 2). На діаграмі розсіювання різними позначками представлено об'єкти, що відносяться до різних кластерів. По осям відкладено значення двох із семи досліджуваних ознак. За станом води у відібраних пробах пункти спостереження розподілилися на 2 групи таким чином: у 2001 році до першого класу увійшли села Тернівка, Мала Лозоватка, до другого — Велика Лозоватка та Завертана, а також північна частина Карачунівського водосховища; у 2007 році в перший кластер виділено село Тернівка, другий містить усі інші об'єкти дослідження.

Такий підхід визначає угрупування пунктів спостереження на певну дату, що дозволяє аналізувати зміни схожості об'єктів у часі. Проте виникає задача визначення об'єктів схожих між собою на всьому часовому проміжку спостереження за всіма досліджуваними показниками одночасно для відображення загальної картини перебігу певних гідрохімічних процесів у воді річки. Для розв'язання такої задачі застосовуємо запропоновану технологію часової кластеризації.

За результатами часової кластеризації було виділено дві групи об'єктів: перша складається з пункту спостереження у с. Тернівка, друга містить усі інші об'єкти дослідження. Таке розбиття на кластери відповідає дійсній гідрологічній та гідрохімічній ситуації на цій ділянці р. Інгулець. До Карачунівського водосховища подається вода каналом «Дніпро–Інгулець». Найбільший вплив цього каналу відзначається на верхній ділянці, що вивчається,

а саме в районі села Тернівка. Нижче за течією вплив дніпровської води на формування хімічного складу води у річці Інгулець менший, більший вплив надає гірничо-збагачувальний комбінат (фільтраційні втрати з гідротехнічних споруд комбінату, пиління хвостосховища та інші).

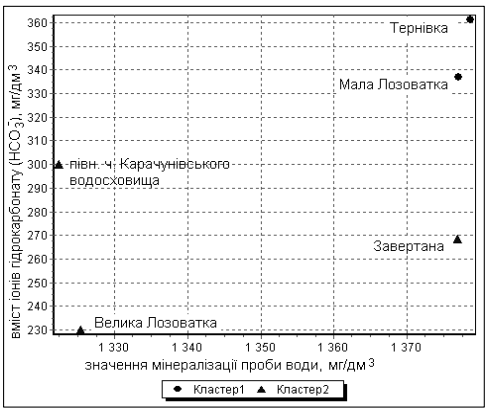

Рис. 1. Діаграма розсіювання. Результати кластеризації за даними 2001 р.

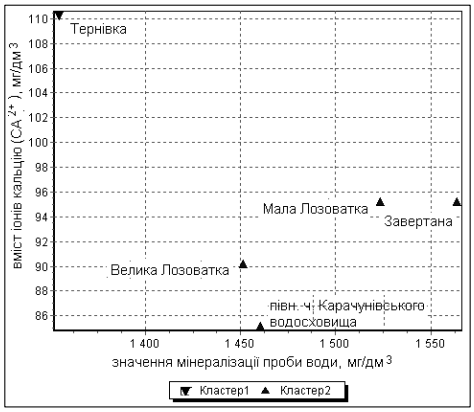

Рис. 2. Діаграма розсіювання. Результати кластеризації за даними 2007 р.

Розглянемо тепер задачу визначення груп об'єктів, які є схожими між собою на досліджуваному часовому проміжку за одним із показників. Тобто вихідні дані $X = \{x_{ijt}\}$, $i = \overline{1,5}$, $j = \overline{1,7}$, $t = \overline{1,9}$ переформуємо у вигляді $X = \{x_{it}\}$, $i = \overline{1,5}$, $t = \overline{1,9}$, де x_{it} — значення досліджуваного показника *i*-го об'єкта у момент часу *t*. Таку задачу можна розв'язати відомими методами кластерного аналізу. На рис. 3–4 представлено діаграми розсіювання об'єктів за значеннями вмісту іонів гідрокарбонату (HCO_3^-) та кальцію (Ca^{2+}) у часовому діапазоні. Отримані результати у більшості випадків співпадають із результатами продемонстрованими запропонованою технологією часової кластеризації, що свідчить про адекватність її результатів. Перевагою запропонованої технології є те, що вона дозволяє виділяти групи схожих об'єктів у часі не за одним обраним показником, а за усіма досліджуваними ознаками.

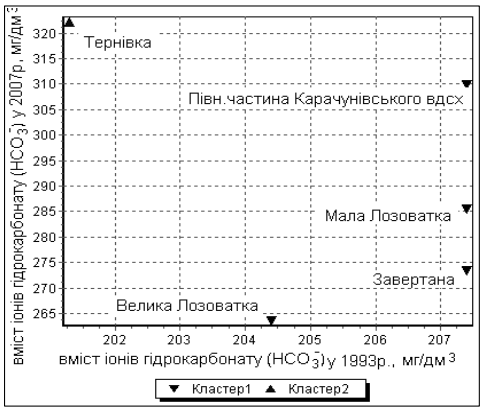

Рис. 3. Результати кластеризації за значеннями вмісту іонів гідрокарбонату (HCO_3^-) у часі

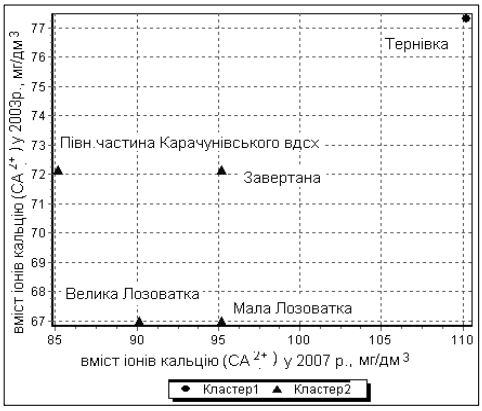

Рис. 4. Результати кластеризації за значеннями вмісту іонів кальцію (Ca^{2+}) у часі

Технологія часової кластеризації увійшла до складу розробленого авторами програмного продукту, що також дозволяє проводити кластерний аналіз даних за певною датою, виділяти групи об'єктів із однорідними значеннями обраного фіксованого показника у багаторічному періоді спостереження, виявляти тенденції та закономірності зміни рівня вмісту головних іонів у воді досліджуваного природного об'єкта за набором ознак у часі.

Такий аналіз дозволяє відобразити загальну картину перебігу певних гідрохімічних процесів у воді річки та визначити часові проміжки зі схожим хімічним складом води для подальшого прийняття рішень щодо планування природоохоронних заходів.

ВИСНОВКИ

У цій роботі запропоновано метод виділення груп об'єктів, схожих між собою за набором ознак, які змінюються у часі, а також технологію багатокритеріальної оцінки якості. Розроблено обчислювальні схеми та створено систему інтелектуального аналізу даних, що реалізує задачі кластеризації, класифікації, візуалізації, обробки та аналізу інформації, забезпечує підтримку прийняття рішень. Наведено та проаналізовано результати практичного застосування розробленої інформаційної технології до даних гідрохімічного моніторингу. Метою аналізу було визначення груп пунктів спостереження, що характеризуються схожим хімічним складом води у р. Інгулець за досліджуваними компонентами, а також виявити тенденції та закономірності зміни вмісту головних іонів у воді досліджуваного природного об'єкта за набором ознак у часі для правильного планування природоохоронних заходів та керування якістю вод річки.

ЛІТЕРАТУРА

1. *Мандель И.Д.* Кластерный аналіз. — М.: Статистика, 1988. — 176 с.
2. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных. — М.: Статистика, 1974. — 240 с.
3. *Jain A.K.* Data clustering: 50 years beyond K-means // Pattern Recognition Letters. — 2010. — 31(8). — P. 651–666.
4. *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: обзор. — М.: Изд. дом НИУ «Высшая школа экономики», 2011. — 88 с.
5. *Бериков В.С., Лбов Г.С.* Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. — 26 с.
6. *Halkidi M., Batistakis Y., Vazirgiannis M.* On Clustering Validation Techniques // Journal of Intelligent Information Systems. — 2011. — **17,** Issue 2–3. — P. 107–145.
7. *Milligan G., Cooper M.* An examination of procedures for determining the number of clusters in a data set // Psychometrika. — 1985. — **50**, № 2. — P. 159–179.
8. *Емельяненко Т.Г., Зберовский А.В., Приставка А.Ф., Собко Б.Е.* Принятие решений в системах мониторинга. — Д.: РИК НГУ, 2005. — 224 с.
9. *Бабак В.П., Білецький А.Я., Приставка О.П., Приставка П.О.* Статистична обробка даних. — К.: МІВВЦ, 2001. — 388 с.
10. *Sarumathi S., Shanthi N., Santhiya G.* A Survey of Cluster Ensemble // International Journal of Computer Applications. — 2013. — **65**, №.9. — P. 8–11.
11. *Бирюков А.С., Рєзанов В.В., Шмаров А.С.* Решение задач кластерного анализа коллективами алгоритмов // Журнал вычислительной математики и математической физики. — 2008. — **48**, № 1. — С. 176–192.

Надійшла 25.05.2012