ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, СИСТЕМНИЙ АНАЛІЗ ТА КЕРУВАННЯ

УДК 519.254

О.Г. Байбуз, д-р техн. наук, проф.,
М.Г.Сидорова

Дніпропетровський національний університет ім. О. Гончара, м.Дніпропетровськ, Україна, e-mail: obaybuz@inbox.ru; Sidorova.m.g@gmail.com

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ БАГАТОВИМІРНИХ ЧАСОВИХ РЯДІВ НА ПРИКЛАДІ ГІДРОХІМІЧНОГО МОНІТОРИНГУ РІЧКИ САМАРА

O.G. Baibuz, Dr. Sci. (Tech.), Professor,
M.G. Sidorova

Oles Honchar Dnipropetrovsk National University, Dnipropetrovsk, Ukraine, e-mail: obaybuz@inbox.ru, Sidorova.m.g@gmail.com

INFORMATION TECHNOLOGY OF THE MULTIVARIATE TIME SERIES FUZZY CLUSTERING ON THE EXAMPLE OF THE SAMARA RIVER HYDROCHEMICAL MONITORING

Мета. Розробка методів для наповнення інформаційної технології нечіткої кластеризації для випадку багатовимірних часових рядів.

Методика. У роботі представлена методика кластерного аналізу багатовимірних часових рядів у вигляді обчислювальної схеми на основі кластеризації одновимірних часових рядів, агрегування результатів у матрицю подібності та визначення на її основі результуючого нечіткого розбиття.

Результати. Адаптовані до кластеризації часових рядів обчислювальні схеми методів: агломеративного ієрархічного, К-середніх, Forel, графового методу найкоротшого незамкненого шляху, що увійшли до ядра запропонованої інформаційної технології. Проведена їх оцінка на основі критеріїв якості. Здійснена практична реалізація до даних гідрохімічного моніторингу техногенно-навантаженої території з аналізом отриманих результатів.

Наукова новизна. Запропонована нова метрика для порівняння часових рядів, яка враховує як характер порівнюваних рядів, так і близькість їх значень, що дозволяє підвищити якість кластеризації. Розроблена інформаційна технологія кластерного аналізу багатовимірних часових рядів на основі ансамблевого підходу та нечіткої логіки.

Практична значимість. На основі запропонованої технології та розробленого програмного забезпечення був проведений кластерний аналіз даних гідрохімічного моніторингу поверхневих вод Західно-Донбаського регіону (р. Самара). Це дозволило виділити групи контрольних створів, що характеризуються схожим фізико-хімічним складом води за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки.

Ключові слова: *кластерний аналіз, часові ряди, міра подібності, інформаційна технологія, гідрохімічний моніторинг*

Постановка проблеми. Аналіз часових рядів і кластеризація є важливими завданнями інтелектуального аналізу даних. В останні роки все більше уваги приділяється об'єднанню цих напрямів, тому що актуальною проблемою є визначення однорідних груп часових рядів для подальшого їх аналізу та прогнозування. Застосування кластеризації часових рядів є корисним у різних прикладних областях: економіці, екології, соціології, маркетингу, медицині та ін.

Актуальною проблемою є аналіз даних гідрохімічного моніторингу. Водні ресурси є одним із найбільш важливих і, разом з тим, найбільш уразливих компонентів довкілля. Особливо складним є гідрохімічний моніторинг водних об'єктів у районах з підвищеним техногенним навантаженням, де екосистема водотоків і водойм знаходиться у важкому стані, ви-

черпуються її природні можливості до самоочищення, існує перспектива якісного вичерпання водних ресурсів. Тому актуальним є проведення гідрохімічного моніторингу з метою збереження, поліпшення та стабілізації якості поверхневих вод для забезпечення оптимальних умов функціонування екосистем, підвищення ефективності природно-господарського комплексу.

Важливу роль у системі гідрохімічного моніторингу відіграє обґрунтування пунктів спостереження, обсягів і періодичності гідрохімічних випробувань. Аналізуючи результати кластерного аналізу, можна певною мірою уніфікувати водоохоронні заходи в межах виділених груп і районів. Визначивши пріоритетні водоохоронні заходи для одного району, планувати й впроваджувати їх для всієї виділеної групи .

Аналіз останніх досліджень та публікацій. Як правило, кластерний аналіз часових рядів складається з п'яти основних етапів: вибір представлення даних, визначення міри схожості порівнюваних рядів, застосування методів кластеризації, визначення якості результатів, аналіз отриманих кластерів.

Існують різні підходи та методи кластеризації, що допускають такі представлення часових рядів: точкове подання (вихідні спостереження); у вигляді послідовності різниці між суміжними значеннями; у вигляді скінченної множини загальних параметрів, таких як показник тренда, сезонності, асиметрії, Херста і т.п.; за допомогою моделей, побудованих на основі вихідних даних [1].

У роботі [2] проведено порівняльний аналіз результатів кластеризації при різних представленнях часових рядів. Кожен з підходів має як переваги, так і недоліки. Наприклад, при великих обсягах даних, щоб прискорити процес кластеризації та позбутися від зашумленности, можна використовувати параметри або моделі часових рядів. Однак при такому підході втрачається частина інформації.

Наступним важливим питанням є визначення близькості між двома часовими рядами. У загальному випадку, ступінь подібності будь-якої пари об'єктів вихідної множини задається або обчисленням відстані між ними на основі деякої метрики, або введенням правила визначення ступеня близькості. Подібність пари часових рядів, як правило, ґрунтується на відстані або кореляції між ними. Найбільш поширеними є евклідова відстань, метрики на основі коефіцієнтів кореляції, DTW (Dynamic time warping) та ін. [3–4]. У випадку кластеризації багатовимірних часових рядів застосування таких метрик призводить до зростання розмірності простору ознак у p разів, що знижує точність кластеризації.

Огляд підходів та методів кластерного аналізу часових рядів проведено в роботах [5–6].

Виділення невирішених раніше частин загальної проблеми. Кластерний аналіз часових рядів є досить актуальною задачею, що підтверджує огляд останніх публікацій. Проте, не зважаючи на активні дослідження, у цій галузі існують невирішені досі проблеми, такі як визначення міри близькості, розробка ефективних алгоритмів кластеризації, аналіз багатовимірних часових рядів.

Задача кластеризації часових рядів гідрохімічного моніторингу поверхневих вод, не зважаючи на свою актуальність, взагалі не знайшла розв'язку.

Постановка задачі. Метою даної роботи є аналіз існуючих підходів та методів кластеризації часових рядів, можливість запропонувати новий спосіб обчислення міри близькості, розробити інформаційну технологію та програмне забезпечення кластеризації багатовимірних часових рядів, провести кластерний аналіз даних гідрохімічного моніторингу.

Нехай маємо множину об'єктів $X = \{x_1, x_2, ..., x_N\}$, що є багатовимірними часовими рядами $x_i = \{u_1, u_2, ..., u_p\}$, $i = \overline{1, N}$; $u_l^{(i)} = \{u_{lt}^{(i)}\}$; $l = \overline{1, p}$, $t = \overline{1, T}$. Кожен об'єкт характеризується p ознаками, значення яких змінюються у часі та досліджуються протягом T моментів спостереження. Усі об'єкти необхідно розподілити на K груп $G_l = \{g_1^{(l)}, g_2^{(l)}, \ldots, g_K^{(l)}\}$; $g_i^{(l)} = \{x_h\}, h = \overline{1, N_i^{(l)}}$; $\sum_{i=1}^{K} N_i^{(l)} = N$; $\bigcup_{i=1}^{K} g_i^{(l)} = X$; $g_i^{(l)} \cap g_j^{(l)} = \varnothing$; $i, j = \overline{1, K}$; $i \neq j$ за схожістю l-ї $l = \overline{1, p}$ досліджуваної ознаки; на основі отриманого набору угрупувань $G = \{G_1, G_2, \ldots, G_p\}$ отримати нечітке розбиття множини X на K груп; проаналізувати отримані результати.

Виклад основного матеріалу. Пропонується інформаційна технологія кластерного аналізу часових рядів, що складається з наступних етапів: вибір та обчислення міри подібності порівнюваних рядів; отримання якісних угрупувань об'єктів аналізу за кожною з досліджуваних ознак; формування агрегованої матриці подібності; отримання результуючого рішення за всіма досліджуваними ознаками; аналіз отриманих кластерів.

Обчислення міри подібності часових рядів. У цій роботі пропонується нова метрика d_{tss} (Time Series Similarity) для визначення подібності часових рядів, що ґрунтується на коефіцієнті кореляції Спірмена та нормованій евклідовій відстані

$$d_{tss} = w_1 \cdot (1 - d_{Sp})/2 + w_2 \cdot d_E^n,$$

де d_{Sp} – коефіцієнт кореляції Спірмена; d_E^n – нормована евклідова відстань; w_1, w_2 – коефіцієнти.

Таким чином, така метрика дозволяє враховувати як характер часових рядів, так і геометричну близькість їх значень.

Слід зауважити, що, оскільки значення $d_{Sp} \in [-1; 1]$, то вираз $(1 - d_{Sp})/2$ змінюється від 0 до 1. Значення нормованої евклідової відстані також належать діапазону від 0 до 1. Коефіцієнти w_1 та w_2 визначають вплив кожної складової, $w_1 + w_2 = 1$.

Для демонстрації переваг запропонованої метрики було проведено ряд експериментів на модельних даних. Результати представлено в табл. 1.

Отримання угрупувань об'єктів аналізу за кожною з досліджуваних ознак. Оскільки не існує універсального методу кластеризації, то до кожного набору даних застосовуємо різні методи кластерного аналізу з метою подальшої оцінки якості отриманих результатів та вибір найкращого рішення. У даній роботі пропонуються обчислювальні схеми на основі адаптації відомих методів до кластеризації часових рядів, а саме: ієрархічних агломеративних, К-середніх, графового методу найкоротшого незамкненого шляху, Forel.

Таблиця 1

Оцінка результатів ієрархічного методу повного зв'язку при виборі різних значеннях w_1 та w_2

Метрика	Індекси якості		
	R	J	FM
d_E^n	0,84	0,46	0,65
d_{Sp}	0,83	0,42	0,61
$d_{tss}, w_1 = 0,1; w_2 = 0,9$	0,85	0,50	0,69
$d_{tss}, w_1 = 0,2; w_2 = 0,8$	0,86	0,51	0,70
$d_{tss}, w_1 = 0,3; w_2 = 0,7$	0,86	0,51	0,70
$d_{tss}, w_1 = 0,4; w_2 = 0,6$	0,86	0,50	0,69
$d_{tss}, w_1 = 0,5; w_2 = 0,5$	0,86	0,50	0,69
$d_{tss}, w_1 = 0,6; w_2 = 0,4$	0,86	0,49	0,67
$d_{tss}, w_1 = 0,7; w_2 = 0,3$	0,85	0,47	0,66
$d_{tss}, w_1 = 0,8; w_2 = 0,2$	0,83	0,43	0,61
$d_{tss}, w_1 = 0,9; w_2 = 0,1$	0,84	0,45	0,64

Обчислювальна схема агломеративного ієрархічного методу:

1. Кожен об'єкт u_i, $i = \overline{1,N}$ вважаємо окремим кластером g_i, $i = \overline{1,K}$, $N_i = 1$. Обираємо метрику та обчислюємо матрицю $D = \{d_{ij}\}, i, j = \overline{1,N}$, де d_{ij} – відстань між часовими рядами u_i та u_j.

2. У матриці відстаней *D* знаходимо мінімальний елемент d_{ij} і кластери g_i та g_j об'єднуємо $g_{i+j} = g_i \cup g_j$, $N_{i+j} = N_i + N_j$.

3. З матриці *D* вилучаємо відстані від g_i та g_j до інших кластерів та додаємо відстані, що відповідають новому кластеру g_{i+j}.

Для обчислення відстані між кластерами існує загальна формула Ланса-Уільямса

$$d(g_{i+j}, g_h) = \alpha_1 d(g_i, g_h) + \alpha_2 d(g_j, g_h) + \beta d(g_i, g_j) + \gamma \left| d(g_i, g_h) - d(g_j, g_h) \right|.$$

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

1) $\alpha_1 = \frac{1}{2}; \alpha_2 = \frac{1}{2}; \beta = 0; \gamma = -\frac{1}{2}$ – одиничного зв'язку (найближчого сусіда);

2) $\alpha_l = \frac{1}{2}; \alpha_m = \frac{1}{2}; \beta = 0; \gamma = \frac{1}{2}$ – повного зв'язку (найвіддаленішого сусіда);

3) $\alpha_1 = \frac{N_i}{N_i + N_j}; \alpha_2 = \frac{N_j}{N_i + N_j}; \beta = 0; \gamma = 0$ – середнього зв'язку;

4) $\alpha_1 = \frac{N_i}{N_i + N_j}; \alpha_2 = \frac{N_j}{N_i + N_j}; \beta = -\frac{N_i N_j}{(N_i + N_j)^2}; \gamma = 0$ – зв'яз-ку між центрами;

5) $\alpha_1 = \frac{N_h + N_i}{N_h + N_i + N_j}; \alpha_2 = \frac{N_h + N_j}{N_h + N_i + N_j}$;

$\beta = -\frac{N_h}{N_h + N_i + N_j}$; $\gamma = 0$ – Уорда.

4. Повторюємо кроки 2–3, доки не отримаємо необхідну кількість кластерів або всі об'єкти не будуть об'єднані в один кластер для побудови дендрограми.

Обчислювальна схема методу К-середніх Болла–Холла:

1. Серед досліджуваних об'єктів u_i, $i = \overline{1,N}$ обираємо деяким чином (випадково, найвіддаленіші, перші) K еталонних u_i^e, $i = \overline{1,K}$, та вважаємо їх рівними центрам окремих кластерів $C_i = (c_{i1}, c_{i2}, ..., c_{iT})$, $i = \overline{1,K}$; $C_i = u_i^e$, $i = \overline{1,K}$. K – кількість кластерів – вхідний параметр алгоритму, $N_i = 1$, $i = \overline{1,K}$.

2. Кожен об'єкт u_i, $i = \overline{1,N}$ відносимо до кластеру g_l

$$d(u_i, C_l) = \min_{j=1,K} d(u_i, C_j); \; N_l = N_l + 1.$$

3. Обчислюємо нові центри кластерів

$$C_i = \frac{1}{N_i} \sum_{j=1}^{N_i} u_j; \quad c_{it} = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{jt}; \quad t = \overline{1,T}.$$

4. Повторюємо кроки 2–3 доки центри не стабілізуються або кількість ітерацій не перевищить задану.

Обчислювальна схема методу К-середніх Мак–Кіна:

1. Серед досліджуваних об'єктів u_i, $i = \overline{1, N}$ обираємо деяким чином (випадково, найвіддаленіші, перші) K еталонних u_i^e, $i = \overline{1, K}$ та вважаємо їх рівними центрам окремих кластерів

$$C_i = (c_{i1}, c_{i2}, ..., c_{iT}), \quad i = \overline{1, K}; \; C_i = u_i^e, \; i = \overline{1, K}.$$

K – кількість кластерів – вхідний параметр алгоритму, $N_i = 1$, $i = \overline{1, K}$.

2. $i = 1$.

3. Об'єкт u_i відносимо до кластеру g_l

$$d(u_i, C_l) = \min_{j=1,K} d(u_i, C_j).$$

4. Обчислюємо новий центр кластера g_l

$$C_l = \frac{N_l \cdot C_l + u_i}{N_l + 1}; \; c_{it} = \frac{N_l \cdot c_{it} + u_{it}}{N_i + 1};$$

$$t = \overline{1, T}; \; N_l = N_l + 1.$$

5. $i = i + 1$. Якщо $i < N$, переходимо на крок 3, інакше – на крок 6.

6. Якщо центри кластерів стабілізувалися, то кінець алгоритму, інакше – переходимо на крок 2.

Обчислювальна схема графового методу найкоротшого незамкненого шляху:

1. Досліджувані об'єкти u_i, $i = \overline{1, N}$ вважатимемо вершинами деякого графа.

2. Обираємо метрику та обчислюємо матрицю $D = \{d_{ij}\}, i, j = \overline{1, N}$, де d_{ij} – відстань між часовими рядами u_i та u_j.

3. У матриці відстаней D знаходимо мінімальний елемент d_{ij}. Вершини графа, що відповідають об'єктам i та j, поєднуємо ребром, довжина якого дорівнює d_{ij}.

4. Доки існують ізольовані вершини, тобто такі, що не мають ребер, знаходимо серед них найближчу до деякої неізольованої точки та утворюємо між ними ребро.

5. Знаходимо у побудованому графі *K-1* найдовших ребра та видаляємо їх. Об'єкти, відповідні вершини яких поєднані ребрами, вважаємо об'єднаними у кластер.

Обчислювальна схема методу FOREL:

1. Вважаємо, що всі об'єкти u_i, $i = \overline{1, N}$ утворюють некластеризовану множину *S*.

2. Випадковим чином обираємо об'єкт $u' \in S$.

3. Утворюємо сферичний кластер g' з центром в точці $C' = u'$ і радіусом R (де R вхідний параметр алгоритму

$$g' := \{u_i \in S \mid d(u_i, C') \le R\}.$$

4. Обчислюємо новий центр кластера g'

$$C' = \frac{1}{N'} \sum_{i=1}^{N'} u_i; \; c_{it} = \frac{1}{N'} \sum_{i=1}^{N'} u_{it}; \; t = \overline{1, T}.$$

5. Повторюємо кроки 3–4, доки центр не стабілізується.

6. Об'єднуємо всі точки сфери з центром C' у кластер та виключаємо з множини *S*.

7. Повторюємо кроки 2–6, доки в *S* є об'єкти.

Оцінка якості. Для оцінки результатів кластеризації існують різноманітні індекси (функціонали, коефіцієнти, показники) якості [7], що дозволяють у кількісному вигляді визначати відповідність вихідного розбиття природній структурі даних (зовнішні критерії), а також порівнювати результати, отримані різними методами або при різних значеннях параметрів (відносні критерії). Запропонована технологія передбачає оцінку якості на основі наступних показників:

- відносні критерії: Калінського-Гарабача, Данна, Беджека-Данна, сума внутрішньокластерних дисперсій за всіма ознаками, сума квадратів відстаней до центрів класів, сума внутрішньокластерних відстаней, відношення середньої внутрішньокластерної та середньої міжкластерної відстаней;

- зовнішні критерії: Ренда, Жакарда, Фолка-Меллоу.

За допомогою зовнішніх критеріїв проведемо оцінку якості результатів кластеризації при виборі різних метрик відстані. Для аналізу було використано набір даних "Synthetic Control Chart Time Series" [4]. Дані представляють собою 600 штучно згенерованих часових рядів 6 типів: нормальні, циклічні, зростаючі, спадаючі, зі здвигом угору, зі здвигом униз.

У табл. 1–2 наведено порівняння отриманих результатів кластеризації в залежності від вибору метрики з правильною кластерною структурою на основі індексів Ренда (R), Жаккарда (J), Фолка-Меллоу (FM).

У табл. 2 наведені результати деяких методів при найкращих комбінаціях значень w_1 та w_2 у порівнянні з евклідовою метрикою та мірою близькості на основі коефіцієнта кореляції Спірмена.

Формування агрегованої матриці подібності. Після визначення угрупувань по кожній ознаці маємо $G_i = \{g_1^{(i)}, g_2^{(i)}, ..., g_K^{(i)}\}$, $i = \overline{1, p}$. Представимо отриману

інформацію у вигляді матриць суміжності A_l, $l \in \overline{1, p}$, а саме $A_l = \left\{a_{ij}^{(l)}\right\}$, $i, j = \overline{1, N}$, де

$$a_{ij}^{(l)} = \begin{cases} 1, \text{ якщо } u_i^{(l)} \in g_k^{(l)} \text{ та } u_j^{(l)} \in g_k^{(l)}, k \in [1, K] \\ 0, \text{ у іншому випадку} \end{cases},$$

тобто $a_{ij}^{(l)} = 1$, якщо часові ряди $u_i^{(l)}$ та $u_j^{(l)}$ відносяться до одного кластеру, $a_{ij}^{(l)} = 0$ у іншому випадку.

Побудуємо агреговану матрицю $A' = \left\{a'_{ij}\right\}$, $i, j = \overline{1, N}$, де $a'_{ij} = \sum_{l=1}^{p} a_{ij}^{(l)} \Big/ p$. Значення a'_{ij} є частотою об'єднання об'єктів x_i та x_j у один кластер при кластеризації за різними ознаками. Чим більше значення a'_{ij}, тим більш подібними є *i*-й та *j*-й об'єкти. Таким чином матрицю A' можна вважати матрицею подібності багатовимірних часових рядів X.

Отримання результуючого рішення. Після того, як були отримані розбиття $G_i = \{g_1^{(i)}, g_2^{(i)}, ..., g_K^{(i)}\}$, $i = \overline{1, p}$ об'єктів аналізу на кластери за кожною з досліджуваних ознак (кластеризовані відповідні одновимірні часові ряди) та визначена агрегована матриця подібності багатовимірних часових рядів A', що характеризує схожість об'єктів за всіма ознаками, необхідно отримати підсумковий розв'язок поставленої задачі, а саме розбиття на кластери об'єктів вихідної множини, що є багатовимірними часовими рядами. 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 розбиття об'єктів на кластери за різними ознаками можуть значно відрізнятися один від одного, то чітка кластеризація багатовимірних часових рядів, тобто кластеризація, при якій кожен об'єкт може відноситися лише до одного з класів, у такому випадку буде малоінформативною. Тому пропонується застосовувати алгоритми кластерного аналізу на основі нечіткої логіки [8]. Це дозволить отримати нечітке розбиття $P = (M^1, ..., M^K)$, $M^l = (\mu_{l1,...,}\mu_{lN})$, де μ_{li} – степінь приналежності *i*-го об'єкта до *l*-го кластера, $0 \le \mu_{li} \le 1$; $\sum_{l=1}^{K} \mu_{li} = 1$; $\sum_{i=1}^{N} \mu_{li} > 0$.

Нечітку кластеризацію будемо проводити методом Уіндхема, який демонструє гарні результати, що відображають не лише матрицю розбиття на класи $P_{K\times N} \in \left[\mu_{li}\right]$, але й матрицю прототипів $V_{K\times N} \in \left[\nu_{li}\right]$, де ν_{li} – вага *i*-го об'єкта як прототипу *l*-го кластера, тобто чим ближче значення ν_{li} до одиниці, тим більше *i*-й об'єкт є прототипом *l*-го кластера, фактично його центром.

Таблиця 2

Порівняння запропонованої метрики на основі оцінки результатів різних методів кластерного аналізу

Метрика	Метод	Індекси якості		
		R	J	FM
d_E^n	Ієрархічн. (ближн. сусіда)	0,88	0,56	0,74
d_{Sp}	Ієрархічн. (ближн. сусіда)	0,18	0,16	0,40
d_{tss}, $w_1 = 0,2; w_2 = 0,8$	Ієрархічн. (ближн. сусіда)	0,89	0,59	0,77
d_E^n	Ієрархічн. (середньої відстані)	0,85	0,49	0,68
d_{Sp}	Ієрархічн. (середньої відстані)	0,69	0,30	0,51
d_{tss}, $w_1 = 0,4; w_2 = 0,6$	Ієрархічн. (середньої відстані)	0,87	0,53	0,71
d_E^n	Ієрархічн. (центр. відстані)	0,82	0,46	0,67
d_{Sp}	Ієрархічн. (центр. відстані)	0,62	0,29	0,53
d_{tss}, $w_1 = 0,3; w_2 = 0,7$	Ієрархічн. (центр. відстані)	0,83	0,50	0,70
d_E^n	Болла-Холла	0,82	0,37	0,54
d_{Sp}	Болла-Холла	0,80	0,35	0,53
d_{tss}, $w_1 = 0,2, w_2 = 0,8$	Болла-Холла	0,88	0,5	0,67
d_E^n	MST	0,88	0,56	0,74
d_{Sp}	MST	0,18	0,16	0,40
d_{tss}, $w_1 = 0,3; w_2 = 0,7$	MST	0,89	0,59	0,77

Метод було адаптовано до кластеризації багатовимірних часових рядів. Далі наведено опис алгоритму та розробленої обчислювальної схеми. Відстань $d(x_i, x_j)$ між об'єктами x_i та x_j будемо розуміти як протиставлення поняттю подібності та обчислювати

наступним чином $d(x_i, x_j) = 1 - a'_{ij}$, оскільки чим більш схожими є об'єкти, тим менша відстань між ними та навпаки.

Алгоритм Уіндхема знаходить розв'язок оптимізаційної задачі

$$Q(P,V) = \sum_{l=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_{li}^2 v_{lj}^2 d(x_i, x_j) \to \min$$

у наступному вигляді

$$P^* = \arg\min_P \begin{Bmatrix} Q(P,V): P = (M^1,...,M^K), \\ M^l = \left[(\mu_{l1,...,}\mu_{lN}),(v_{l1,...,}v_{lN})\right] \end{Bmatrix};$$

$$0 \le \mu_{li} \le 1;\ 0 \le v_{lj} \le 1;\ \sum_{l=1}^{K} \mu_{li} = 1;\ \sum_{j=1}^{N} v_{li} = 1;$$

$$\sum_{i=1}^{N} \mu_{li} > 0;\ \sum_{l=1}^{K} v_{lj} > 0;\ i, j = 1,...,N;\ l = 1,...,K.$$

Обчислювальна схема методу Уіндхема:

1. Випадковим чином задаємо матрицю розбиття $P_{K\times N} \in \left[\mu_{li}\right];\ 0 \le \mu_{li} \le 1;\ \sum_{i=1}^{N} \mu_{li} > 0$. Таким чином отримуємо початкове розбиття $P_{(0)} = (M_{(0)}^1,...,M_{(0)}^K)$ на K нечітких кластерів.

2. $i = 1$.

3. Покладаємо $P_{(i)} = P_{(0)}$ і обчислюємо матрицю прототипів $V_{(0)}$ наступним чином

$$v_{li} = \frac{1 \Big/ \sum_{i=1}^{n} \mu_{li}^2 d(x_i, x_j)}{\sum_{j=1}^{n} (1 \Big/ \sum_{i=1}^{n} \mu_{li}^2 d(x_i, x_j))}; \qquad (1)$$

$i, j = 1,...,N;\ l = 1,...,K$.

4. Покладаємо $V_{(i+1)} = V_{(i)}$ і обчислюємо матрицю розбиття $P_{(i)}$ наступним чином

$$\mu_{li} = \frac{1 \Big/ \sum_{j=1}^{n} v_{lj}^2 d(x_i, x_j)}{\sum_{k=1}^{c} (1 \Big/ \sum_{j=1}^{n} v_{kj}^2 d(x_i, x_j))}; \qquad (2)$$

$i, j = 1,...,N;\ l = 1,...,K$.

5. Обчислюємо матрицю прототипів $V_{(i)}$ на основі $P_{(i)}$ у відповідності зі співвідношенням (1).

6. Якщо $Q(P_i,V_i) - Q(P_{i-1},V_{i-1}) < \varepsilon$, то $P^* = P_{(i)}$, $V^* = V_{(i)}$, інакше $i = i + 1$ і переходимо на крок 4.

При фіксованій матриці прототипів V матриця розбиття P, що будується у відповідності до формули (2), мінімізує $Q(P_i,V)$ по всім матрицям розбиття P_i, так що виконується $Q(P_i,V_{i-1}) \le Q(P_{i-1},V_{i-1})$. Аналогічно $Q(P_i,V)$ мінімізується матрицею V, що будується за формулою (1), так що $Q(P_i,V_i) \le Q(P_i,V_{i-1})$. Таким чином, алгоритм зменшує значення цільової функції $Q(P,V)$ на кожній ітерації. Враховуючи це, а також невід'ємність функції $Q(P,V)$, можна зробити висновок, що послідовність $\{Q(P_i,V_i)\}$ є збіжною, при будь-якому $\varepsilon > 0$ алгоритм зупиниться після скінченної кількості кроків.

Практична реалізація та аналіз отриманих результатів. Запропонована технологія була застосована до даних гідрохімічного моніторингу Західно-Донбаського регіону (р.Самара, Дніпропетровська область, Україна). Метою роботи було визначення груп пунктів спостереження, що характеризуються схожим хімічним складом води у р.Самара за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки.

Проби води відбиралися у 5 пунктах спостереження: с.Коханівка (у межах населеного пункту), с.Маломиколаївка (6 км вище населеного пункту, 1км вище скиду шахтних вод зі ставка-накопичувача б.Косьмінна), с.Петрівка (у межах населеного пункту, 1км нижче скиду шахтних вод зі ставка-накопичувача б.Косьмінна), с.Богуслав (у межах населеного пункту), с.Тернівка (в/з Федора, 2км нижче населеного пункту, 1км вище скиду шахтних вод зі ставка-накопичувача б.Свідовок), с.Вербки (у межах населеного пункту, 1км нижче скиду шахтних вод з б.Свідовок, 1км вище скиду з очисних споруд комунального підприємства м.Павлоград), м.Павлоград, с.В'язовок (у межах населеного пункту, 1км нижче скиду з очисних споруд комунального підприємства м.Павлоград), с.Кочережки (у межах населеного пункту). Для кожної проби фізико-хімічними методами аналізу визначалися наступні показники: водневий показник (РН), розчинений у воді кисень (O_2), біохімічне споживання кисню (БСК), хімічне споживання кисню (ХСК), нітрати (NO_3), нітрити (NO_2), фосфати (PO_4), сухий залишок (СЗ), завислі речовини (ЗР), хлориди (Cl), сульфати (SO_4), аміак (NH_4), нафтопродукти (НП) протягом наступних моментів часу: 04.04.2000, 20.06.2000, 02.10.2000, 27.03.2001, 06.06.2001, 03.09.2001, 12.03.2002, 30.09.2002, 22.04.2003, 23.06.2003, 23.09.2003,

22.03.2004, 23.06.2004, 20.09.2004, 23.03.2005, 29.06.2005, 20.09.2005.

Таким чином об'єктами $X = \{x_1, x_2, ..., x_9\}$ є контрольні створи, $u_l^{(i)} = \{u_{lt}^{(i)}\}$, $l = \overline{1,13}$, $t = \overline{1,17}$ – значення l-го досліджуваного показника, що вимірюється в і-й пробі води протягом контрольних моментів часу.

За даними, що характеризують часові зміни кожної окремої ознаки, було проведено кластерний аналіз різними методами. На основі критеріїв якості визначені найкращі розбиття. Результати приналежності кожного об'єкта певному кластеру за відповідною ознакою наведені в табл. 3. Для визначення кількості кластерів було застосовано аналіз дендрограми та критерій Калінського-Гарабача [9]. На рис. 1–3 представлені результати дослідження оптимальної кількості кластерів за даними, що відповідають показнику РН. Як бачимо, обидва методи, а також візуальний аналіз дендрограми, свідчать про те, що найкращі результати кластеризації досягаються при двох кластерах. Подібні результати були отримані й для інших ознак.

У табл. 4 наведено результати нечіткої кластеризації об'єктів аналізу за всіма досліджуваними показниками.

Контрольні створи було розподілено на два нечітких кластера наступним чином: до першого кластеру належать с.Коханівка, с.Богуслав, с.Тернівка, с.Вербки, м.Павлоград, с.Кочережки; другий кластер склали с.Маломиколаївка, с.Петрівка та с.В'язовок.

Таблиця 3

Результати кластерного аналізу за гідрохімічними показниками

Пункти спостереження	Гідрохімічні показники												
	РН	O_2	БСК	ХСК	NO_3	NO_2	PO_4	СЗ	ЗР	Cl	SO_4	NH_4	НП
Коханівка	1	1	1	1	1	1	1	1	1	1	1	1	1
Маломиколаївка	2	1	2	1	1	1	1	2	1	1	2	1	1
Петрівка	2	2	1	1	1	1	1	2	1	2	2	1	1
Богуслав	2	2	1	2	1	1	1	1	1	1	2	1	1
Тернівка	2	1	1	1	1	1	1	2	1	1	2	1	1
Вербки	2	1	1	1	1	1	1	2	1	2	2	1	1
Павлоград	2	2	2	1	2	2	2	1	1	1	1	2	2
В'язовок	1	1	1	1	1	1	1	1	1	1	1	1	1
Кочережки	1	1	2	1	1	1	1	2	2	1	2	1	1

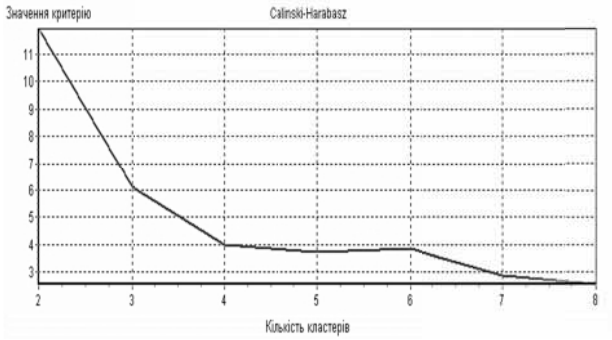

Рис. 1. Дослідження оптимальної кількості кластерів за критерієм Калінського-Гарабача

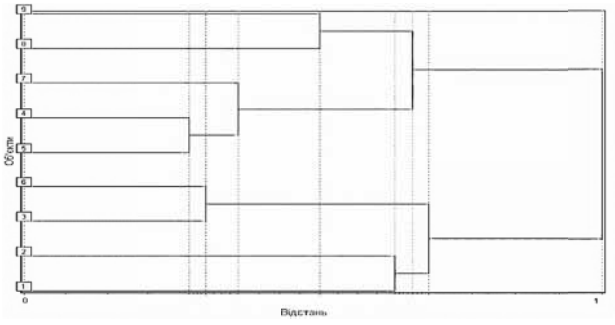

Рис. 2. Дендрограма (процес об'єднання об'єктів у кластери)

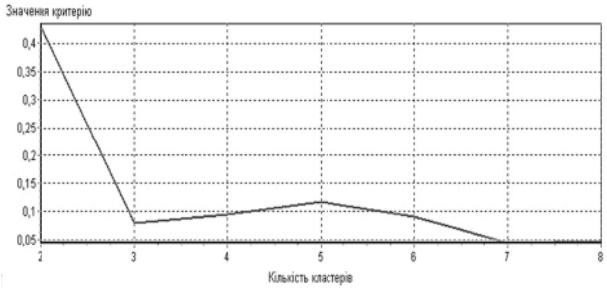

Рис. 3. Дослідження оптимальної кількості кластерів за критерієм різниці між рівнями об'єднання на дендрограмі

Таблиця 4

Результати нечіткої кластеризації вихідних даних

Пункти спостереження	Міра приналежності	
	Кластер 1	Кластер 2
Коханівка	0,5008	0,4992
Маломиколаївка	0,4984	0,5016
Петрівка	0,4989	0,5011
Богуслав	0,5002	0,4998
Тернівка	0,5008	0,4992
Вербки	0,5016	0,4984
Павлоград	0,5002	0,4998
В'язовок	0,4988	0,5012
Кочережки	0,5003	0,4997

Висновки. У роботі розроблена інформаційна технологія кластерного аналізу багатовимірних часових рядів на основі ансамблевого підходу та нечіткої логіки; запропонована нова метрика для порівняння часових рядів, що враховує як характер порівнюваних рядів, так і близькість їх значень; наведені та проаналізовані результати практичного застосування розробленої технології до даних гідрохімічного моніторингу. Була вирішена практична задача визначення груп пунктів спостереження, що характеризуються схожим хімічним складом води у р.Самара за досліджуваними компонентами, а також виявлення схожих тенденцій і закономірностей зміни вмісту физико-хімічних показників у воді досліджуваного природного об'єкта для правильного планування природоохоронних заходів та керування якістю вод річки. Перспективами подальших розробок є підвищення якості отримуваних результатів шляхом удосконалення запропонованої інформаційної технології за рахунок розробки нових методів та залучення експертних оцінок, а також проведення більш детального аналізу отриманих результатів на основі методів статистичної обробки даних.

Список літератури / References

1. Wang, X., Smith, K., Hyndman, R.and Alahakoon, D. (2001), "A scalable method for time series clustering", *Tech. Report Department of Econometrics and Business Statistics at Monash University*, Melbourne, Australia.

2. Паршутин С.В. Кластеризация временных рядов с применением карт самоорганизации: сборник научных трудов / С.В. Паршутин // Интегрированные модели и мягкие вычисления в искусственном интеллекте. – Коломна. – 2007. – С. 465–472.

Parshutin, S.V. (2007), "Time series clustering with application of Self-Organizing Maps", *Int. Conf. Proc. "Integrated Models and Soft Computing in Artificial Intelligence"*, Kolomna, pp. 465–472.

3. Iglesias, F. and Kastner, W., (2013), "Analysis of similarity measures in times series clustering for the discovery of building energy patterns", *Energies*, Vol. 6, pp. 579–597.

4. Alcock, R.J. and Manolopoulos, Y., (1999), "Time-Series Similarity Queries Employing a Feature-Based Approach.", *7th Hellenic Conference on Informatics*. August 27–29, Ioannina, Greece.

5. Liao, T.W., (2005), "Clustering of time series data – survey", *Pattern Recognition*, Vol. 38, pp. 1857–1874.

6. Rani, S., Sikka, G., (2012), "Recent Techniques of Clustering of Time Series Data: A Survey", *International Journal of Computer Applications*, Vol. 32, no. 15, pp. 1–9.

7. Гусарова Л. Проверка обоснованности кластерного решения / Л. Гусарова, И. Яцкив // Reliability and statistics in transportation and communication (RelStat). – 2004. –Т. 5. – № 2. – С. 49–56.

Gusarova, L. and Yatskiv, I., (2003),"Checking of the cluster solution validity", *Proc. of Int. Conf. RelStat*, Vol. 5, no. 2, pp. 49–56.

8. Вятченин Д.А. Нечеткие методы автоматической классификации: монография / Вятченин Д.А. – Минск: УП „Технопринт", 2004. – 219с.

Vyatchenin, D.A., (2004), *Nechetkie metody avtomaticheskoy klassifikatsii* [Fuzzy Methods of Automatic Classification], Monograph, *Tehnoprint*, Minsk, Belarus.

9. Яцкив И. Методы определения количества кластеров при классификации без обучения / И. Яцкив, Л. Гусарова // Transport and Telecommunication. – 2003. – Т. 4. – № 1. – С. 23–28.

Yatskiv, I. and Gusarova, L., (2003), "Methods for determining the number of clusters in unsupervised classification", *Transport and Telecommunication*, Vol. 4, no. 1, pp. 23–28.

Цель. Разработка методов для наполнения информационной технологии нечеткой кластеризации в случае многомерных временных рядов.

Методика. В работе представлена методика кластерного анализа многомерных временных рядов в виде вычислительной схемы на основе кластеризации одномерных временных рядов, агрегирования результатов в матрицу сходства и определения на ее основе результирующего нечеткого разбиения.

Результаты. Адаптировано к кластеризации временных рядов вычислительные схемы методов: агломеративного иерархического, К-средних, Forel, графового метода кратчайшего незамкнутого пути, которые вошли в ядро предложенной информационной технологии. Проведена их оценка на основе критериев качества. Осуществлена практическая реализация к данным гидрохимического мониторинга техногенно-нагруженной территории с анализом полученных результатов.

Научная новизна. Предложена новая метрика для сравнения временных рядов, которая учитывает как характер сравниваемых рядов, так и близость их значений, что позволяет повысить качество кластеризации. Разработана информационная технология кластерного анализа многомерных временных рядов на основе ансамблевого подхода и нечеткой логики.

Практическая значимость. На основе предложенной технологии и разработанного программного обеспечения был проведен кластерный анализ данных гидрохимического мониторинга поверхностных вод Западно-Донбасского региона (р.Самара). Это позволило выделить группы контрольных створов, характеризующихся похожим физико-химическим составом воды по исследуемым компонентам для правильного планирования природоохранных мероприятий и управления качеством вод реки.

Ключевые слова: *кластерный анализ, временные ряды, мера сходства, информационная технология, гидрохимический мониторинг*

Purpose. Development of the methods for filling the information technology of fuzzy clustering in the case of multivariate time series.

Methodology. This paper presents a technique of cluster analysis of multivariate time series as a computa-

tional schemes based on the one-dimensional time series clustering, aggregating results into a similarity matrix and determination the result fuzzy partition.

Findings. Computational schemes of methods: agglomerative hierarchical, K-means, Forel, graph method of the shortest non-closed path have been adapted to the time series clustering and included in the core of the proposed information technology. Their quality has been assessed by different quality criteria. The practical implementation with the analysis of the results has been applied to the data of hydrochemical monitoring of technologically-laden area.

Originality. A new metric for comparing time series which takes into account both the nature of the compared series and the closeness of their values has been proposed, that can improve the quality of clustering. The information technology of multivariate time series clustering based on an ensemble approach and fuzzy logic has been proposed.

Practical value. On the basis of the proposed technology and developed software cluster analysis of data hydrochemical monitoring of surface waters of the West Donbass region (r. Samara) has been held. It has allowed to identify groups of control points, which characterized by similar physical and chemical composition of water on the investigated components for proper environmental planning and management of water quality of the river.

Keywords: *cluster analysis, time series, measure of similarity, information technology, hydro-chemical monitoring*

Рекомендовано до публікації докт. техн. наук О.М. Карповим. Дата надходження рукопису 10.10.13.