УДК 004.891.2:550.8.052

Кондратенко Н. Р.[1], Снігур О. О.[2]

[1]Канд. техн. наук, доцент, професор кафедри захисту інформації Вінницького національного технічного університету, Вінниця, Україна

[2]Аспірант кафедри захисту інформації Вінницького національного технічного університету, Вінниця, Україна

ІНТЕРВАЛЬНИЙ НЕЧІТКИЙ КЛАСТЕРНИЙ АНАЛІЗ ДЛЯ МОНІТОРИНГУ СТАНУ АРТЕЗІАНСЬКОЇ СВЕРДЛОВИНИ

Актуальність. Моніторинг природних систем різного характеру є необхідною умовою раціонального природокористування. Технології інтелектуального аналізу даних, зокрема кластерний аналіз, надають широкі можливості для візуалізації наборів даних, що дозволяє використовувати ці технології людьми, які не мають спеціальної математичної підготовки. Задача моніторингу системи, стан якої змінюється в часі, висуває вимогу розширеної інтерпретації результатів кластеризації з урахуванням історичних даних. Технічні можливості для виявлення характеру змін, що відбуваються в об'єкті, представленому набором даних, мають особливе значення в задачі моніторингу водних ресурсів, оскільки вони перебувають у тісному взаємозв'язку з зовнішнім середовищем, та величина їхніх запасів залежить від багатьох факторів, зовнішніх відносно водоносної системи. Після введення в експлуатацію артезіанська свердловина потребує постійного спостереження задля правильного керування експлуатацією підземних вод, захисту їх від забруднення та вичерпання, а також попередження негативних наслідків впливу водовідбору на навколишнє середовище. Крім того, для складних природних систем характерна висока надлишковість простору параметрів, а також наявність як відомих, так і не виявлених досі кореляційних зв'язків між параметрами. Ці фактори зумовлюють необхідність використання методів кластерного аналізу, здатних працювати в умовах невизначеності та надлишковості параметрів.

Мета роботи – розширення можливостей для аналізу зміни стану системи в часі шляхом урахування невизначеностей, присутніх у даних спостережень.

Метод. Запропоновано застосування методу інтервального нечіткого кластерного аналізу для дослідження зміни характеристик набору даних у часі та виявлення загальних тенденцій. Формалізація поставленої технологічної задачі в термінах інтелектуального аналізу даних передбачає можливість одночасної роботи з множиною вхідних векторів. Сформульовано покроковий алгоритм побудови інтервальної оцінки стану природної системи на основі історичних даних спостережень та поточних значень.

Результати. Запропоновану модель адаптовано до розв'язання технологічної задачі моніторингу артезіанської свердловини та експериментально показано можливості раннього виявлення прихованих закономірностей.

Висновки. Інтервальний нечіткий кластерний аналіз дозволяє враховувати та моделювати невизначеності довільної природи, що виникають у даних досліджень артезіанської свердловини на різних стадіях моніторингу. Показано, що одночасне подання на вхід системи даних кількох свердловин може дати змогу оцінити не лише їхнє розташування щодо стандартних компактних класів за (потенційною) якістю води, але й взаємне розташування, і в кінцевому підсумку вказати на деяку не виявлену до цього закономірність.

Ключові слова: кластерний аналіз, інтервальні ступені належності, інтервальні нечіткі множини, критерії якості кластеризації, візуалізація даних.

НОМЕНКЛАТУРА

$SC(c, m)$ – індекс розбиття (Partition Index);
$K(c, m)$ – критерій Квона;
$XB(c, m)$ – критерій Хіе-Бені;
μ_{ij} – ступінь належності точки j до кластера i;
ν_j – центр j-го кластера;
m – рівень нечіткості;
C – кількість кластерів;
N – кількість точок;
$\overline{\nu}$ – середнє значення центрів кластерів;
(X, Y) – набір даних спостережень;
W^i – артезіанська свердловина;
$p_1 ... p_m$ – простір ознак (параметрів свердловини);
$X^i = \{x_1^i, ..., x_m^i\}$ – результати дослідження свердловини W^i за параметрами $p_1 ... p_m$;
$Y = \{y_1, ..., y_n\}$ – класи, утворені об'єктами x^i відповідно до оцінюваного параметра.

ВСТУП

Методи та моделі нечіткого кластерного аналізу мають широке поле для застосування в сучасних інтелектуальних системах. У контексті технологій Data Mining одним із основних призначень кластеризації є наочне подання (візуалізація) результатів обчислень, що дозволяє використовувати ці технології людьми, які не мають спеціальної математичної підготовки [1]. Кластерний аналіз широко використовується для виділення прихованих закономірностей та внутрішніх взаємозв'язків у великих масивах багатовимірних даних, таких як обробка зображень, розпізнавання образів, дослідження та прогнозування соціально-економічних процесів тощо. Одним із важливих його застосувань є також попередня обробка наборів даних, зокрема виділення інформативних ознак при роботі з надлишковими даними.

У системах, стан яких змінюється в часі, виникають більш широкі можливості для інтерпретації результатів кластеризації. В умовах змінних значень параметрів об'єктів видається перспективним дослідження зміни характеру розбиття видозміненого набору даних у порівнянні з вихідним. Прикладом застосування такого підходу можуть бути системи моніторингу об'єктів та процесів різного походження, як природних, так і технічних. В рамках дослідження цей підхід буде застосовано в задачі моніторингу артезіанської свердловини. Після введення в експлуатацію вона потребує постійного спостереження задля правильного керування експлуатацією підземних вод, захисту їх від забруднення та вичерпання, а також попередження негативних наслідків впливу во-

DOI 10.15588/1607-3274-2017-4-9

довідбору на навколишнє середовище [2]. Завдання поточного та періодичного моніторингу свердловини в багатьох випадках покладається на організацію, що здійснювала роботу зі свердловиною від початку гідрогеологічної розвідки, пов'язаної з даним проектом. З одного боку це означає, що в розпорядженні дослідника є всі дані попередніх спостережень; з іншого, якщо у віданні організації знаходиться все родовище або ж його частина, що охоплює значну кількість свердловин, детальний аналіз кожного з контрольованих параметрів на предмет потенційно небезпечних відхилень вимагає суттєвих затрат часу. Якщо взяти до уваги надлишковість простору параметрів, що завжди характерна для складних природних систем, та відомі кореляційні зв'язки між ними, задача ускладнюється ще більше. Крім того, в системі можуть також існувати досі не виявлені зв'язки між параметрами, які не можуть бути враховані експертом-людиною.

Кластерний аналіз як інструмент подання, або візуалізації, даних дозволяє виявити приховані закономірності та внутрішні взаємозв'язки, присутні в досліджуваному наборі даних [3–7]. Серед існуючих на сьогоднішній день методів кластеризації є й такі, що показують добрі результати на даних високої розмірності [3, 4, 7–9]. Об'єктом даного дослідження є методи інтервальної нечіткої кластеризації на основі альтернативних критеріїв якості [6]. Предметом дослідження є можливості застосування цього методу для розв'язання задачі моніторингу природних систем. Мета роботи – розширення можливостей для аналізу зміни стану системи в часі шляхом урахування невизначеностей, присутніх у даних спостережень. Задля досягнення поставленої мети вирішуються такі задачі:

– формалізувати задачу моніторингу природної системи в термінах Data Mining;

– сформулювати алгоритм побудови інтервальної оцінки стану артезіанської свердловини на основі попередніх та поточних даних спостережень;

– показати можливість роботи методу на даних спостережень природних систем.

1 ПОСТАНОВКА ЗАДАЧІ

Задачу моніторингу однієї або більше артезіанських свердловин в термінах кластерного аналізу можна сформулювати таким чином. Нехай задано набір даних спостережень:

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_m^1 \\ x_1^2 & x_2^2 & & x_m^2 \\ \dots & & & \\ x_1^n & x_2^n & & x_m^n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}.$$

Оцінюваний параметр, в розрізі якого проводиться дослідження, умовно назвемо «перспективність свердловини». В даному випадку це узагальнена кількісна оцінка стану свердловини, що відображає перспективність її подальшої експлуатації, а також ступінь невизначеності, пов'язаної з цією оцінкою. Для набору (*X, Y*), який у подальшому будемо називати навчальним набором, відомо висновок експерта про належність свердловини W^i до одного з класів за перспективністю. Відомо також, що об'єкти $x^1, \dots, x^n$ утворюють компактні кластери в просторі вхідних ознак. На множину *X* також накладається умова репрезентативності її відносно генеральної сукупності векторів ознак, тобто множина *X* повинна містити представників усіх *c* класів. Необхідно розбити множину *X* на *c* кластерів та визначити ступені належності до кожного з *c* кластерів довільної свердловини W^z, що описується вхідним вектором $X^z = \left\{ x_1^z, \dots, x_m^z \right\}$, $X^z \notin X$.

Як буде показано далі, така постановка допускає можливість одночасного розв'язання задачі не лише для одного вхідного вектора, а й для матриці, побудованої з даних досліджень множини свердловин.

2 ОГЛЯД ЛІТЕРАТУРИ

В роботі [10] здійснено спроби розв'язання задач регіонального районування та соціально-економічного прогнозування. Роботи [11, 12, 13] демонструють, як кластерний аналіз може виконувати сегментацію множини абонентів провайдера телекомунікаційних послуг. Проте математичні методи, що лежать в основі цих досліджень, суттєво обмежені припущенням, що вхідні дані є абсолютно точними, правдивими та незашумленими. Метод, запропонований в роботах [14, 15], попри високі оптимізаційні властивості, ставить аналогічну вимогу. Відомо, що на практиці такі умови трапляються вкрай рідко, тому дана задача вимагає методів кластерного аналізу, стійких до викидів та шуму. Метод PCM (Possibilistic C-Means) [16] задовольняє цій вимозі – він надзвичайно стійкий до шумів у вхідних показниках, але базується на нечітких множинах типу 1. Це не дає змогу дати повністю адекватну оцінку досліджуваній множині даних, оскільки крім точок, що вносять шум, у характеристиках кожної точки закладена певна невизначеність, яка не може не перенестись на результат кластеризації. При цьому характеризувати ступінь належності точки до кластера одним числом недостатньо. Зважаючи на це, подання ступенів належності у вигляді інтервальних значень та застосування математичного апарату нечітких множин типу 2 в задачі кластеризації має практичний сенс.

Методи нечіткої кластеризації дають також позитивні результати в задачі загального оцінювання якості води [17, 18]. Робота з підземними водами ускладнюється їхньою недоступністю для безпосередніх спостережень. Інформація про стан системи достеменно відома в окремих точках родовища; дані ж про інші ділянки отримують, екстраполюючи фактичні точкові дані на ділянки, про які фактичної інформації немає [19]. Тому сучасні методи та технології оцінювання якості підземних вод [20–22] в цілому суттєво не відрізняються від методів, що застосовуються для досліджень поверхневих вод. З усіх факторів, що впливають на якість та особливості видобутку підземних вод особливу увагу приділено антропогенному забрудненню [23, 24] та дослідженню вразливості водоносних горизонтів до шкідливих речовин, присутніх у повітрі, ґрунтах та поверхневих водах [25, 26]. Математичні моделі та методи, які при цьому застосовуються, не передбачають моделювання невизначеностей, що виникають при спостереженні гідрогеологічних систем. Всі вони побудовані на припущенні, що отримані дані спостережень точні, повністю визначені, однозначні та достовірні. У випадку такого складного об'єкта спос-

тережень як підземні води задовольнити ці вимоги до вибірки даних практично неможливо, що не може не вплинути на адекватність побудованих моделей. Врахування та моделювання невизначеностей, закладених у вихідному наборі даних, у ряді випадків дає змогу помітити тенденції та зміни в характері процесів, що протікають у природній системі, на стадії їх формування [27].

Розв'язання задачі моніторингу підземних вод на основі методу інетрвального нечіткого кластерного аналізу дасть змогу підвищити ефективність спостережень та скерувати увагу спеціалістів на можливі негативні фактори й тенденції на ранніх фазах їх розвитку. Широкі можливості для одночасної роботи з множиною свердловин дозволять підвищити частоту таких контролюючих заходів.

3 МАТЕРІАЛИ І МЕТОДИ

Для розв'язання задачі поточного моніторингу артезіанської свердловини застосуємо модель на основі модифікованого методу кластеризації РСМ з інтервальним виходом, запропоновану в роботі [6]. Він має за основу метод можливісної кластеризації [16] та передбачає отримання інтервальних значень ступенів належності об'єктів до кластерів за рахунок регулювання рівня нечіткості. Інтервал зміни рівня нечіткості визначається за допомогою критеріїв якості кластеризації Квона, Хіе-Бені та індексу розбиття [5]:

$$SC(c,m)=\sum_{i=1}^{c}\frac{\sum_{k=1}^{N}\left(\mu_{i,k}\right)^m\left\|x_k-\nu_i\right\|^2}{\sum_{k=1}^{N}\mu_{i,k}\sum_{j=1}^{c}\left\|\nu_j-\nu_i\right\|^2}, \quad (1)$$

$$K(c,m)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^m\left\|x_j-\nu_i\right\|^2+\frac{1}{c}\sum_{i=1}^{c}\left\|\nu_i-\bar{\nu}\right\|^2}{\min\limits_{i\neq j}\left\|\nu_i-\nu_j\right\|^2}, \quad (2)$$

$$XB(c,m)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^m\left\|x_j-\nu_i\right\|^2}{N\min\limits_{i,j}\left\|\nu_i-x_j\right\|^2}. \quad (3)$$

Процес прийняття рішення відбувається в такій послідовності.

1. Визначити параметри розбиття навчального набору (X,Y) на кластери відповідно до методу [6]. В даному його застосуванні суттєвими є лише значення рівня нечіткості та координати центрів кластерів. Обчислювати остаточні значення ступенів належності зразків навчального набору до отриманих кластерів немає необхідності.

2. На останньому кроці методу [6] обчислити значення ступенів належності до кластерів для зразка, що являє собою вектор W^z параметрів контрольованої свердловини. Обчислення відбувається за формулою (4) методу РСМ:

$$\mu_{ij}=\frac{1}{1+\left(\frac{d_{ij}^2}{\eta_j}\right)^{\frac{1}{m-1}}}. \quad (4)$$

3. В загальному випадку векторів може бути більш ніж один; в такому випадку йдеться про «підміну» навчального набору даних тестовим. Оскільки для оцінювання позиції кожного з контрольованих зразків достатньо лише обчислити ступені належності за формулою (4), на обсяг тестового набору даних, що може оброблятися, не накладається обмежень; в загальному випадку він може перевищувати обсяг навчального набору.

4. Обчислити значення критерію (критеріїв) якості кластеризації на об'єднанні навчального та тестового наборів даних. Пропонується використати критерії (1–3), значення яких досліджуються на кроках 1–2 протягом виконання інтервальної кластеризації навчального набору даних. Це дасть змогу оцінити, чи спотворюють дані тестового набору «ідеальне» розбиття, отримане для навчального набору, та визначити кількісну міру цих спотворень.

5. Остаточне рішення приймається за ступенями належності точки, що характеризується вектором X^z, до кожного з c утворених кластерів.

4 ЕКСПЕРИМЕНТИ

Об'єктом кластеризації будемо вважати набір значень параметрів артезіанської свердловини $X^i=\left\{x_1^i,...,x_m^i\right\}$, включаючи такі, що описують особливості геологічної будови, тектонічні, кліматичні та гідрогеологічні умови, а також результати дослідних робіт безпосередньо в свердловині: дані геофізичних досліджень, пробних і дослідних відкачок, параметри, що характеризують якість підземних вод. Кластерний аналіз відбувається в просторі ознак свердловини x_1, ... x_{84}, приклади яких наведено в табл. 1.

Навчальний набір даних побудовано на основі архівних даних досліджень свердловин родовищ підземних вод, розташованих на території Правобережної Геологічної Експедиції. Вхідному вектору, що містить усі параметри свердловини x_1–x_{84}, ставиться у відповідність висновок експерта-гідрогеолога про її придатність до видобутку питної води терміном на найближчі 5 років. Навчальний набір даних складається з 20 зразків, приклади зразків наведено в табл. 2.

Таблиця 1 – Параметри гідрогеологічного дослідження

Позначення змінної	Назва параметру	Область значень
x_1	Віддаленість від населених пунктів, км	0–50
x_2	Віддаленість від шосейних доріг загальнодержавного значення, км	0–50
...		
x_{83}	Гідрогеологічні умови за ступенем складності	1–3
x_{84}	Гідрогеологічні умови за ступенем вивченості	0–10

В ході експерименту досліджено можливості запропонованого вище методу, визначивши параметри розбиття на основі навчального набору даних з табл. 2 та застосувавши їх до тестового набору, частково показаного в табл. 3.

Тестовий набір даних складається з 30 зразків, що не входять до навчального, та імітує множину вхідних даних задачі моніторингу реальних артезіанських свердловин.

5 РЕЗУЛЬТАТИ

Розіб'ємо навчальний набір даних на кластери за методом [6]. Число кластерів вважаємо наперед заданим, c=3.

Поведінку критеріїв якості розбиття відносно рівня нечіткості показано на рис. 1.

За правилом, запропонованим в [6], інтервальне значення рівня нечіткості

$$\tilde{m} = \tilde{m}_K \cup \tilde{m}_{XB} \cup \tilde{m}_{SC} = [1{,}7;\ 3{,}5] \cup [1{,}7;\ 3{,}5] \cup [1{,}6;\ 4{,}1].$$

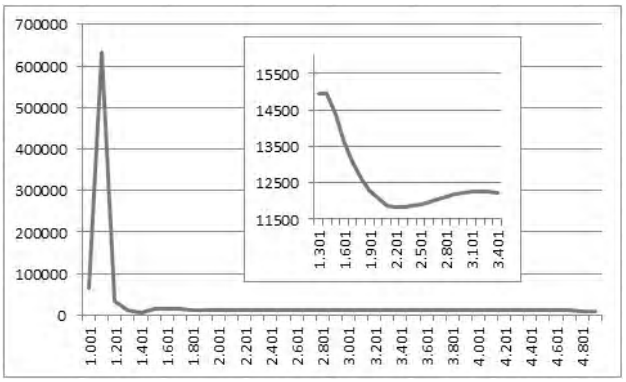
Побудувавши розбиття для правої та лівої границь інтервалу $\tilde{m}$, отримано центри кластерів та ступені належності зразків навчального набору даних, наведені в таблицях 4–5.

Аналіз розташування центрів та складу кластерів в розрізі поняття «перспективність свердловини» дозволяє поставити у відповідність кластерам значення перспективності: кластер 1 – висока, кластер 2 – недостатня, кластер 3 – достатня.

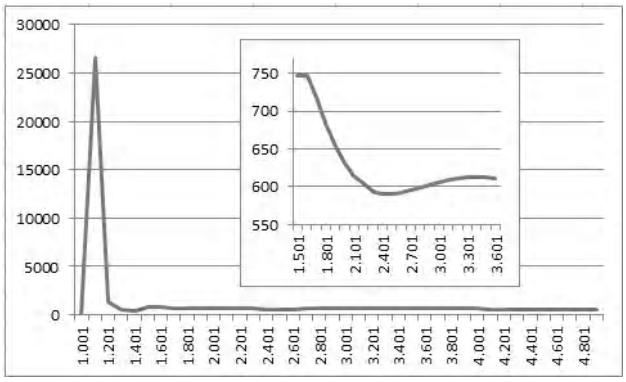
Значення критеріїв якості кластеризації на навчальному наборі даних становлять $SC(c,m)$ = 155; $K(c,m)$ = 11880,2; $XB(c,m)$ = 589,9.

Отримані значення рівня нечіткості та координат центрів кластерів приймаємо за вихідні для обчислення ступенів належності зразків із тестової вибірки. Запропонований метод не передбачає повторного обчислення центрів кластерів та знаходження оптимального рівня нечіткості, тому час виконання обчислень лінійно залежить від кількості зразків у тестовому наборі. Результати обчислень ступенів належності зразків тестового набору даних до трьох кластерів подано в табл. 6.

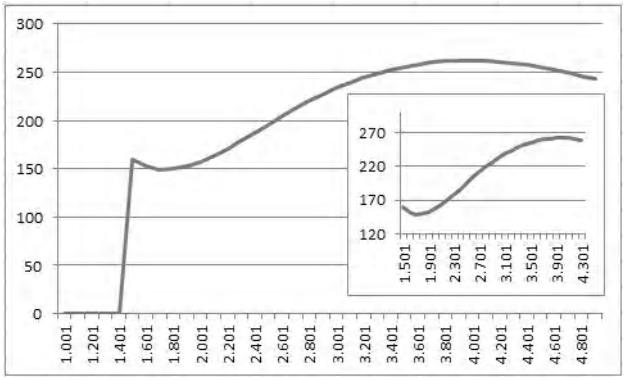
Значення критеріїв якості кластеризації на тестовому наборі даних становлять $SC(c,m)$ = 218,9; $K(c,m)$ = 12034,4; $XB(c,m)$ = 626,5.



а



б



в

Рисунок 1 – Поведінка критеріїв якості відносно рівня нечіткості:
а – критерій Квона; б – критерій Хіе-Бені; в – індекс розбиття

Таблиця 2 – Дані досліджень свердловин (навчальна вибірка)

Змінна	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_1	1,5	3	12	22	18	15	15	26	37	35	27	4	48	50	50	30	25	22	39	31
x_2	43	12	12,5	25	4	11	32	2	5	2	10	3	3	0,8	4,5	1	40	14	1,5	3
...																				
x_{83}	2	2	2	2	3	2	2	3	2	2	3	3	2	2	3	3	2	2	2	2
x_{84}	10	7	7	10	10	10	7	10	7	7	7	7	7	7	7	7	7	7	7	7

Таблиця 3 – Тестовий набір даних

Змінна	1	2	3	4	5	6	7	8	9	10	…	26	27	28	29	30
x_1	30	25	22	39	37	35	27	12	22	18		26	37	35	27	39
x_2	1	40	14	1,5	5	2	10	12,5	25	4		2	5	2	10	1,5
...																
x_{83}	2	2	3	3	2	3	2	2	2	3		3	2	2	3	3
x_{84}	7	7	7	10	7	10	10	10	10	10		10	7	7	7	10

Таблиця 4 – Центри кластерів

Змінна	Кластер 1		Кластер 2		Кластер 3	
	Ліва границя	Права границя	Ліва границя	Права границя	Ліва границя	Права границя
x_1	25,82	29,13	20,98	20,99	18,23	24,11
x_2	34,17	37,36	12,91	16	14,14	22,87
...						
x_{83}	0,9	0,99	1,05	2,82	1,48	2,24
x_{84}	7,98	9,88	5,06	9,6	6,3	6,33

Таблиця 5 – Ступені належності

№	Кластер 1		Кластер 2		Кластер 3	
	Ліва границя	Права границя	Ліва границя	Права границя	Ліва границя	Права границя
1	0,694	0,858	0,107	0,468	0,196	0,536
2	0,064	0,594	0,286	0,494	0,227	0,362
…						
19	0,148	0,867	0,351	0,387	0,076	0,332
20	0,209	0,730	0,167	0,382	0,323	0,494

Таблиця 6 – Ступені належності зразків тестової вибірки

№	Кластер 1		Кластер 2		Кластер 3		Результат / ширина інтервалу	Оцінка експерта
	Ліва границя	Права границя	Ліва границя	Права границя	Ліва границя	Права границя		
1	0,165	0,178	0,618	0,918	0,249	0,491	Недостатня / 0,3	Недостатня
…								
4	0,722	0,99	0,029	0,326	0,053	0,352	Висока / 0,27	Висока
…								
30	0,661	0,96	0,211	0,389	0,119	0,422	**Висока / 0,3**	**Недостатня**

6 ОБГОВОРЕННЯ

Змістовність отриманих значень критеріїв якості кластеризації та характер їх інтерпретації визначається природою тестової вибірки в кожному конкретному випадку. В рамках даного дослідження тестова вибірка складалася з даних довільних свердловин на різних стадіях експлуатації, тому складно простежити закономірність, яка могла б пояснити розходження в значеннях критеріїв. Загальні критерії якості характеризують тестову вибірку в цілому та мають практичний сенс лише тоді, коли вона формувалася за деяким принципом або системою. Можливі застосування запропонованого методу, в яких значення критеріїв якості після об'єднання навчальної та тестової вибірки можуть нести змістове навантаження, доступне для інтерпретації в термінах предметної галузі. Наприклад, тестова вибірка може складатися з даних різних свердловин одного родовища в заданий момент часу, скажімо, за 10 років після введення їх в експлуатацію. В цьому випадку незначне розходження в значеннях критеріїв якості до та після внесення тестового набору може говорити про стабільність гідрогеологічної системи та процесів, що в ній відбуваються, і широкі перспективи подальшої експлуатації артезіанських свердловин на досліджуваній території.

У даному ж випадку слід звернути увагу на розходження між індивідуальними значеннями інтервалів ступенів належності зразків тестової вибірки до кластерів. Дослідження моделі на основі інтервальної кластеризації зокрема показують розходження рішення, прийнятого системою, з експертним висновком у прикладі 30 (табл. 6). Зразок 30 за всіма показниками крім одного (концентрація радону, 219 Бк/дм3) близький до кластера 1. Оскільки кластерний аналіз як технологія навчання без учителя не має можливостей для врахування інших факторів, окрім Евклідової відстані між точками в просторі ознак, зразок 30 віднесено до кластера 1 (Висока), хоча насправді вода з такими характеристиками непридатна до вживання. Безперечно, якщо розглядати зразок 30 окремо від даних інших спостережень, результат роботи системи в цьому випадку слід вважати незадовільним. В будь-якому разі, всі системи підтримки прийняття рішень у галузі моделювання гідрогеологічних процесів вимагають коригування за допомогою експертних знань. Проте якщо зважити на те, що зразок 30 за своїми значеннями практично повністю повторює зразок 4, а також на те, що зразок 30 імітує поступове виникнення негативної тенденції в часі (підвищення радіоактивності), то до оцінювання результату слід підходити по-іншому. В даному випадку спостерігається розширення зони невизначеності ([0,661; 0,96]) в порівнянні з попереднім значенням ([0,722; 0,99]) та зсув ступеня належності до «хорошого» кластера в бік зменшення. Водночас помітна зміна ступенів належності зразка до інших кластерів у бік зростання: [0,211; 0,389] порівняно з [0,029; 0,326] для зразка №4; [0,119; 0,422] проти [0,053; 0,352] відповідно.

Таким чином, у контексті задачі моніторингу артезіанської свердловини результат, отриманий для зразка 30 достатньо змістовний для того, щоб звернути увагу дослідника на процеси, що відбуваються в цій свердловині, та вказує на необхідність більш детального дослідження. В решті випадків результат роботи системи повністю уз-

годжується з рішенням експерта для відповідного зразка; ширину інтервалу можна вважати мірою невизначеності, спричиненої браком експертних знань. Вона досить суттєва, як і слід очікувати від такого складного об'єкта дослідження як гідрогеологічна система.

ВИСНОВКИ

Роботу присвячено розширенню галузі застосування методів кластерного аналізу шляхом аналізу зміни характеру розбиття набору даних у часі. Подання ступенів належності в інтервальній формі робить можливим вивчення цих змін та їх якісну інтерпретацію. Інтервальний нечіткий кластерний аналіз дозволяє враховувати та моделювати невизначеності довільної природи, що виникають в об'єкті спостереження на різних стадіях моніторингу. Ця властивість має особливу цінність у задачах моніторингу водних ресурсів, оскільки вони перебувають у тісному взаємозв'язку з зовнішнім середовищем, та величина їхніх запасів залежить від багатьох факторів, зовнішніх відносно водоносної системи. Тому регулярний моніторинг є невід'ємною складовою процесу дослідження та експлуатації артезіанських свердловин і дає можливість виявити ранні ознаки вичерпання джерела водопостачання, а також зміни в складі вод або глибини залягання водоносного горизонту.

В рамках дослідження виконано адаптацію методу інтервального нечіткого кластерного аналізу до прикладної задачі моніторингу стану підземних вод. Формалізація поставленої технологічної задачі в термінах інтелектуального аналізу даних передбачає можливість одночасної роботи з множиною вхідних векторів. Сформульовано покроковий алгоритм побудови інтервальної оцінки стану свердловини на основі історичних даних спостережень та поточних значень. Подання ступенів належності в інтервальній формі дозволяє враховувати та моделювати невизначеності, пов'язані з браком експертних знань. Останнє має особливо важливе значення в контексті кластерного аналізу як технології навчання без учителя. Отримані результати перевірено експериментально. Показано, що одночасне подання на вхід системи даних кількох свердловин може дати змогу оцінити не лише їхнє розташування щодо стандартних компактних класів за (потенційною) якістю води, але й взаємне розташування, і в кінцевому підсумку вказати на деяку не виявлену до цього закономірність.

СПИСОК ЛІТЕРАТУРИ

1. Дюк В. А. Data Mining: учебный курс / В. А. Дюк, А. П. Самойленко. – СПб. : Изд. Питер, 2001. – 368 с.
2. Петровська М. А. Охорона вод (санітарні норми і правила): навч. посібник / М. А. Петровська. – Львів : Видавничий центр Львівського національного університету імені Івана Франка, 2005. – 205 с.
3. Субботин С. А. Выделение набора информативных признаков на основе эволюционного поиска с кластеризацией / С. А. Субботин, А. А. Олейник // Штучний інтелект. – 2008. – № 4. – С. 704–711.
4. Cai W. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation / W. Cai, S. Chen, D. Zhang // Pattern Recognition. – 2007. – Vol. 40, № 3. – P. 825–838.
5. Oliveira J. V. Advances in Fuzzy Clustering and Its Applications / J. V. Oliveira, W. Pedrycz. – Sidney : John Wiley & Sons, 2007. – 435 p.
6. Кондратенко Н. Р. Інтервальна нечітка кластеризація на основі альтернативних критеріїв якості / Н. Р. Кондратенко, О. О. Снігур // Наукові вісті НТУУ «КПІ». – 2012. – № 2. – С. 59–66.
7. Martyniuk T. B. Formalization of the Object Classification Algorithm / T. B. Martyniuk, A. V. Kozhemiako, L. M. Kupershtein // Cybernetics and Systems Analysis. – 2015. – Vol. 51, № 5. – P. 751–756.
8. Bankruptcy forecasting: a hybrid approach using fuzzy c- means clustering and multivariate adaptive regression splines (MARS) / [J. De Andres, P. Lorca, F. J. D. C. Juez et al.] // Expert Systems with Applications. – 2011. – № 38. – P. 1866–1875.
9. A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints / [J. Wang, J. Kong, Y. Lu et al.] // Computerized Medical Imaging and Graphics. – 2008. – Vol. 32, № 8. – P. 685–698.
10. Зайченко Ю. П. Нечеткие модели и методы в интеллектуальных системах / Ю. П. Зайченко. – К. : «Издательский дом «Слово», 2008. – 344 с.
11. Захарченко С. М. Використання генетичного алгоритму в задачі кластеризації абонентів інтернет-провайдера / С. М. Захарченко, Н. Р. Кондратенко, О. О. Манаєва // Інформаційні технології та комп'ютерна інженерія : І Міжнародна науково-практична конференція, Вінниця, 19–21 травня 2010 р. : тези доповідей. – Вінниця : ВНТУ, 2010. – С. 120–121.
12. Захарченко С. М. Дослідження можливостей генетичного алгоритмув задачі кластеризації користувачів мережі Інтернет / С. М. Захарченко, Н. Р. Кондратенко, О. О. Манаєва // Інформаційні технології та комп'ютерна інженерія. – Вінниця : ВНТУ. – 2010. – № 2 (18). – С. 68–72.
13. Манаєва О. О. Побудова кластерів з використанням генетичного алгоритму / О.О. Манаєва // XXXIX науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області, Вінниця, 10–12 березня 2010 р. : тези доповідей. – Вінниця : ВНТУ, 2010.
14. Кондратенко Н. Р. Нечітка кластеризація абонентів інтернет-провайдера / Н. Р. Кондратенко, О. О. Манаєва // Наукові праці Вінницького національного технічного університету. – 2011. – № 2.
15. Кондратенко Н. Р. Нечітка кластеризація з урахуванням індексу вірогідності в задачах соціального спрямування / Н. Р. Кондратенко, О. О. Манаєва // Системний аналіз та інформаційні технології: матеріали Міжнародної науково-технічної конференції SAIT 2011. – К. : ННК «ІПСА» НТУУ «КПІ». – 2011. – С. 265.
16. Krishnapuram R. A Possibilistic Approach to Clustering / R. Krishnapuram, J. M. Keller // IEEE Transactions on Fuzzy Systems. – 1993. – № 1 (2). – P. 98–110.
17. A fuzzy technique for food- and water quality assessment with an electronic tongue / [B. Iliev, M. Lindquist, L. Robertsson et al.] // Fuzzy Sets and Systems – 2006. – Vol. 157, №. 9. – P. 1155–1168.
18. Assessment of the surface water quality in Northern Greece / [V. Simeonov, J. A. Stratis, C. Samara et al.] // Water Research. – 2003. – Vol. 37, № 17. – P. 4119–4124.
19. Боревский Б. В. Оценка запасов подземных вод / Б. В. Боревский, Н. И. Дробноход, Л. С. Язвин. – 2-е изд., перераб. и доп. – К. : Выща шк. Головное изд-во, 1989. – 407 с.
20. Analysis of groundwater quality using fuzzy synthetic evaluation / [S. Dahiya, B. Singh, S. Gaur et al.] // Journal of Hazardous Materials. – 2007. – Vol. 147, № 3. – P. 938–946.
21. Use of fuzzy synthetic evaluation for assessment of groundwater quality for drinking usage: a case study of Southern Haryana, India / [B. Singh, S. Dahiya, S. Jain, et al.] // Environmental Geology. – 2008. – Vol. 54, № 2. – P. 249–255.

22. Transient Ground-Water Flow Simulation Using a Fuzzy Set Approach / [C. Dou, W. Woldt, M. Dahab et al.] // Groundwater. – 2005. – Vol. 35, № 2. – P. 205–215.
23. An integrated fuzzy-stochastic modeling approach for risk assessment of groundwater contamination / [J. Li, G. H. Huang, G. Zeng et al.] // Journal of Environmental Management. – 2007. – Vol. 82, № 2. – P. 173–188.
24. Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool / [R. C. M. Nobre, O. C. Rotunno Filho, W. J. Mansur et al.] // Journal of Contaminant Hydrology. – 2007. – Vol. 97, № 3. – P. 277–292.
25. Dixon B. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis / B. Dixon // Journal of Hydrology. – 2005. – Vol. 309, № 1. – P. 17–38.
26. Dixon B. Groundwater vulnerability mapping: A GIS and fuzzy rule based integrated tool / B. Dixon // Applied Geography. – 2005. – Vol. 25, № 4. – P. 327–347.
27. Kondratenko N. Interval Fuzzy Modeling of Complex Systems under Conditions of Input Data Uncertainty / N. Kondratenko, O. Snihur // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 4/4 (82). – P. 20–28.

Стаття надійшла до редакції 14.03.2017.
Після доробки 17.05.2017.

Кондратенко Н. Р.[1], Снигур О. А.[2]
[1]Канд. техн. наук, доцент, профессор кафедры защиты информации Винницкого национального технического университета, Винница, Украина
[2]Аспирант кафедры защиты информации Винницкого национального технического университета, Винница, Украина

ИНТЕРВАЛЬНЫЙ НЕЧЕТКИЙ КЛАСТЕРНЫЙ АНАЛИЗ ДЛЯ МОНИТОРИНГА СОСТОЯНИЯ АРТЕЗИАНСКОЙ СКВАЖИНЫ

Актуальность. Мониторинг природных систем различного характера является необходимым условием рационального природопользования. Технологии интеллектуального анализа данных, в частности кластерный анализ, предоставляют широкие возможности для визуализации наборов данных, что позволяет использование этих технологии людьми, не имеющими специальной математической подготовки. Задача мониторинга системы, состояние которой изменяется во времени, выдвигает требование расширенной интерпретации результатов кластеризации с учетом исторических данных. Технические возможности для выявления характера изменений, происходящих в объекте, представленном набором данных, имеют особое значение в задаче мониторинга водных ресурсов, поскольку они находятся в тесной взаимосвязи с внешней средой, и величина их запасов зависит от многих факторов, внешних по отношению к водоносной системе. После введения в эксплуатацию артезианская скважина нуждается в постоянном наблюдении для правильного управления эксплуатацией подземных вод, защиты их от загрязнения и истощения, а также предупреждения негативных последствий влияния водоотбора на окружающую среду. Кроме того, для сложных природных систем характерна высокая избыточность пространства параметров, а также наличие как известных, так и не выявленных ранее корреляционных связей между параметрами. Эти факторы обуславливают необходимость использования методов кластерного анализа, способных работать в условиях неопределенности и избыточности параметров.

Цель работы – расширение возможностей для анализа изменения состояния системы во времени путем учета неопределенностей, присутствующих в данных наблюдений.

Метод. Предложено применение метода интервального нечеткого кластерного анализа для исследования изменения характеристик набора данных во времени и выявления общих тенденций. Формализация поставленной технологической задачи в терминах интеллектуального анализа данных предусматривает возможность одновременной работы с множеством входных векторов. Сформулирован пошаговый алгоритм построения интервальной оценки состояния природной системы на основе исторических данных наблюдений и текущих значений.

Результаты. Предложенная модель адаптирована к решению технологической задачи мониторинга артезианской скважины. Экспериментально показаны возможности раннего выявления скрытых закономерностей.

Выводы. Интервальный нечеткий кластерный анализ позволяет учитывать и моделировать неопределенности произвольной природы, возникающих в данных исследований артезианской скважины на разных стадиях мониторинга. Показано, что одновременная подача на вход системы данных нескольких скважин может позволить оценить не только их расположение относительно стандартных компактных классов по (потенциально) качеству воды, но и их взаимное расположение, и в конечном итоге указать на некоторую не обнаруженную ранее закономерность.

Ключевые слова: кластерный анализ, интервальные степени принадлежности, интервальные нечеткие множества, критерии качества кластеризации, визуализация данных.

Kondratenko N. R.[1], Snihur O.O.[2]
[1]PhD, Associate professor, Professor of department of information security, Vinnytsia National Technical University, Vinnytsia, Ukraine
[2]Postgraduate student of department of information security, Vinnytsia National Technical University, Vinnytsia, Ukraine

INTERVAL FUZZY CLUSTER ANALYSIS FOR ARTESIAN WEL L STATE MONITORING

Context. Monitoring natural systems of diverse nature is an essential condition of rational environmental management. Data Mining technologies, cluster analysis in particular, provide a wide range of capabilities for data sets visualization, which makes it possible for these technologies to be used by individuals with no specialized background in mathematics. The task of monitoring a system that changes its state in time requires extended interpretation of clustering result, which would allow accounting for historical data. Technical capabilities for revealing the nature of changes occurring in the object represented by a data set are of particular importance in water resources monitoring area, as they are strongly related to their environment, and the quantity of the available reserves depend on multiple factors, which are external to the water-bearing system. Upon commissioning, an artesian well requires constant monitoring in order to ensure proper management of groundwater processing, protection against pollution and exhaustion, and preventing negative effects of groundwater mining on the environment. In addition, high redundancy of the parameter space is typical for complex natural systems, as well as existence of both known and not yet discovered correlations between parameters. These factors necessitate the use of cluster analysis methods, which would be capable of operating within the conditions of uncertainty and parameter redundancy.

Objective. The goal of the research is extending the capabilities for analyzing changes in a system's state in time by accounting for uncertainties present in observations data.

Method. An application of the interval fuzzy cluster analysis method for investigating changes in data set characteristics in time, and for revealing general trends, is proposed. Formalizing the technological problem faced by the research in terms of Data Mining provides for a

possibility of simultaneously processing multiple input vectors. A step-by-step algorithm for interval evaluation of the state of a natural system based on historical observations data and current values is developed.

Results. The proposed model is adapted for solving the technological task of an artesian well monitoring, and its capabilities for revealing hidden patterns on early stages are demonstrated experimentally.

Conclusions. Interval fuzzy cluster analysis allows taking into account and modeling uncertainties of any given nature, which may occur in artesian well research data on different stages of monitoring. It is shown, that concurrent input of multiple wells data may allow to evaluate not only there position against the standard compact classes according to (potential) water quality, but also their position against each other, and eventually indicate a previously unknown pattern.

Keywords: cluster analysis, interval membership grades, interval fuzzy sets, clustering validity indices, data visualization.

REFERENCES

1. Dyuk V. A., Samoylenko A. P. Data Mining: uchebnyiy kurs. Sankt-Peterburg, Izd. Piter, 2001, 368 p.
2. Petrovska M. A. Okhorona vod (sanitarni normy i pravyla): Navch. Posibnyk. Lviv, Vydavnychyi tsentr Lvivskoho natsionalnoho universytetu imeni Ivana Franka, 2005, 205 p.
3. Subbotin S. A., Oleynik A. A. Vyidelenie nabora informativnyih priznakov na osnove evolyutsionnogo poiska s klasterizatsiey, *Shtuchniy Intelekt*, 2008, No. 4, pp. 704–711.
4. Cai W., Chen S., Zhang D. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation, *Pattern Recognition*, 2007, Vol. 40, No. 3, pp. 825–838.
5. Oliveira J. V., Pedrycz W. Advances in Fuzzy Clustering and Its Applications. Sidney, John Wiley & Sons, 2007, 435 p.
6. Kondratenko N. R., Snihur O. O. Intervalna nechitka klasteryzatsiia na osnovi alternatyvnykh kryteriiv yakosti, Naukovi visti NTUU «KPI», 2012, No. 2, pp. 59–66.
7. Martyniuk T. B., Kozhemiako A. V., Kupershtein L. M. Formalization of the Object Classification Algorithm, *Cybernetics and Systems Analysis*, 2015, Vol. 51, No. 5, pp. 751–756.
8. Andres J. De, Lorca P., Juez F. J. D. C. et al. Bankruptcy forecasting: a hybrid approach using fuzzy c- means clustering and multivariate adaptive regression splines (MARS), *Expert Systems with Applications*, 2011, No. 38, pp. 1866–1875.
9. Wang J., Kong J., Lu Y. et al. A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints, *Computerized Medical Imaging and Graphics,* 2008, Vol. 32, No. 8, pp. 685–698.
10. Zaychenko Yu. P. Nechetkie modeli i metodyi v intellektualnyih sistemah. Kiev, «Izdatelskiy dom «Slovo», 2008, 344 p.
11. Zakharchenko S. M., Kondratenko N. R., Manaieva O. O. Vykorystannia henetychnoho alhorytmu v zadachi klasteryzatsii abonentiv internet-provaidera, *Informatsiini tekhnolohii ta kompiuterna inzheneriia : I Mizhnarodna naukovo-praktychna konferentsiia, Vinnytsia, 19–21 travnia 2010 r. : tezy dopovidei*. Vinnytsia, VNTU, 2010, pp. 120–121.
12. Zakharchenko S. M., Kondratenko N. R., Manaieva O. O. Doslidzhennia mozhlyvostei henetychnoho alhorytmuv zadachi klasteryzatsii korystuvachiv merezhi Internet, *Informatsiini tekhnolohii ta kompiuterna inzheneriia*. Vinnytsia, VNTU, 2010, No. 2 (18), pp. 68–72.
13. Manaieva O. O. Pobudova klasteriv z vykorystanniam henetychnoho alhorytmu, *KHIKh naukovo-tekhnichna konferentsiia profesorsko-vykladatskoho skladu, spivrobitnykiv ta studentiv universytetu z uchastiu pratsivnykiv naukovo-doslidnykh orhanizatsii ta inzhenerno-tekhnichnykh pratsivnykiv pidpryiemstv m. Vinnytsi ta oblasti, Vinnytsia, 10–12 bereznia 2010 r. : tezy dopovidei.* Vinnytsia: VNTU, 2010.
14. Kondratenko N. R., Manaieva O. O. Nechitka klasteryzatsiia abonentiv internet-provaidera, *Naukovi pratsi Vinnytskoho natsionalnoho tekhnichnoho universytetu*, 2011, No. 2.
15. Kondratenko N. R., Manaieva O. O. Nechitka klasteryzatsiia z urakhuvanniam indeksu virohidnosti v zadachakh sotsialnoho spriamuvannia, *Systemnyi analiz ta informatsiini tekhnolohii: materialy Mizhnarodnoi naukovo-tekhnichnoi konferentsii SAIT 2011*, Kiev, NNK «IPSA» NTUU «KPI», 2011, P. 265.
16. Krishnapuram R., Keller J. M. A Possibilistic Approach to Clustering, *IEEE Transactions on Fuzzy Systems*, 1993, No. 1 (2), pp. 98–110.
17. Iliev B., Lindquist M., Robertsson L. et al.A fuzzy technique for food- and water quality assessment with an electronic tongue, *Fuzzy Sets and Systems*, 2006, Vol. 157, No. 9, pp. 1155–1168.
18. Simeonov V., Stratis J. A., Samara C. et al. Assessment of the surface water quality in Northern Greece, *Water Research*, 2003, Vol. 37, No. 17, pp. 4119–4124.
19. Borevskiy B. V., Drobnohod N. I., Yazvin L. S. Otsenka zapasov podzemnyih vod, 2-e izd., pererab. i dop. Kiev, Vyischa shk. Golovnoe izd-vo, 1989, 407 p.
20. Dahiya S., Singh B., Gaur S. et al. Analysis of groundwater quality using fuzzy synthetic evaluation, *Journal of Hazardous Materials*, 2007, Vol. 147, No. 3, pp. 938–946.
21. Singh B., Dahiya S., Jain S., et al. Use of fuzzy synthetic evaluation for assessment of groundwater quality for drinking usage: a case study of Southern Haryana, India, *Environmental Geology*, 2008, Vol. 54, No. 2, pp. 249–255.
22. Dou C., Woldt W., Dahab M. et al.Transient Ground-Water Flow Simulation Using a Fuzzy Set Approach, *Groundwater*, 2005, Vol. 35, No. 2, pp. 205–215.
23. Li J., Huang G. H., Zeng G. et al. An integrated fuzzy-stochastic modeling approach for risk assessment of groundwater contamination, *Journal of Environmental Management*, 2007, Vol. 82, No. 2, pp. 173–188.
24. Nobre R. C. M., Rotunno Filho O. C., Mansur W. J. et al.Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool, *Journal of Contaminant Hydrology*, 2007. Vol. 97, No. 3, pp. 277–292.
25. Dixon B. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis, *Journal of Hydrology*, 2005, Vol. 309, No. 1, pp. 17–38.
26. Dixon B. Groundwater vulnerability mapping: A GIS and fuzzy rule based integrated tool, *Applied Geography*, 2005, Vol. 25, No. 4, pp. 327–347.
27. Kondratenko N., Snihur O. Interval Fuzzy Modeling of Complex Systems under Conditions of Input Data Uncertainty, *Eastern-European Journal of Enterprise Technologies*, 2016, Vol. 4/4 (82), pp. 20–28.