ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

УДК 004.8

Н. Р. Кондратенко[1]
О. О. Снігур[1]

ЕВОЛЮЦІЙНИЙ ПОШУК ІНФОРМАТИВНИХ ОЗНАК ІЗ ЗАЛУЧЕННЯМ ЕКСПЕРТА В ЗАДАЧІ ОЦІНКИ ЯКОСТІ АРТЕЗІАНСЬКОЇ ВОДИ

[1]Вінницький національний технічний університет

Розв'язано задачу виділення набору інформативних ознак, що описують стан артезіанської свердловини. Запропоновано метод автоматичного вибору комбінації ознак, який інтегрує використання емпіричних спостережень та експертних знань і дозволяє скоротити простір ознак за рахунок таких, що не чинять впливу на значення вихідного параметра, або впливом яких можна знехтувати.

Ключові слова: система підтримки прийняття рішень, експертна система, інтелектуальний аналіз даних, кластеризація, виділення інформативних ознак, гідрогеологія, еволюційний пошук.

Вступ

На сьогодні існує два основні підходи до побудови систем підтримки прийняття рішень. Перший — на основі експериментальних даних, в яких система навчається на основі емпіричної інформації. Другим традиційним підходом є експертні системи, що базуються на експертних знаннях, що дозволяють легко вводити та використовувати знання, формалізовані експертами, однак не володіють можливостями навчання в процесі побудови та експлуатації за спостереженнями.

Тому в задачах інтелектуального аналізу багатомірних даних гостро стоїть задача інтеграції експертних знань та емпіричної інформації в одній системі, яка враховувала б, з одного боку, досвід експерта, а з іншого — експериментальні дані, накопичені на основі спостережень за реальними об'єктами [1]. Основною складністю при цьому є виникнення надлишковості двох видів: надлишковість правил (два та більше правила, що несуть однакову інформацію) та надлишковість ознак (ускладнення антецедентів правил побудови нечіткого логічного висновку за рахунок нерелевантних ознак). Відкидання надлишкових правил — задача відносно тривіальна, та в рамках цього дослідження не розв'язується. У цій роботі розв'язується задача спрощення простору ознак з метою виділення серед них мінімального набору інформативних.

Кількість ознак, що використовуються для описування досліджуваних об'єктів, нарівні з якісним складом множини ознак, чинить значний вплив на кінцевий результат аналітичних процедур. Використання надлишкових та недостатньо інформативних ознак в процесі аналізу даних не лише веде до зростання обчислювальної складності, але й негативно позначається на якості роботи алгоритмів прийняття рішень, оскільки деякі ознаки можуть вносити шум, в той час як дві або більше інших несуть якісно однакову інформацію про об'єкт.

Задача автоматичного визначення набору інформативних ознак має особливе значення в галузях, що вимагають роботи з великими обсягами багатомірних даних. Наприклад, для оцінки якості води, що видобувається з артезіанської свердловини, виникає необхідність роботи з набором вхідних даних надзвичайно високої розмірності, кожен вхідний вектор якого містить понад 300 ознак [2].

У цій роботі розв'язується задача визначення такої підмножини вихідного набору ознак, яка б містила лише найбільш інформативні ознаки, що описують заданий об'єкт, для подальшого аналізу даних на основі цієї підмножини. У цьому випадку отриманий набір ознак пропонується використовувати для опису входів нечіткої логічної системи-класифікатора. Прикладною галуззю застосування результатів дослідження є оцінка якості артезіанської води. Класифікатор може на основі отриманого набору ознак відносити вектор, що подається на вхід, до одного з класів за якістю.

Наразі відома низка методів визначення інформативних ознак: методи на основі повного пере-

бору, методи послідовного додавання та видалення ознак, ранжування ознак [3]. Однак такі методи пов'язані з необхідністю комбінаторного перебору, що робить їх мало застосовними на практиці, або ж вони використовують критерії оцінювання індивідуальної інформативності ознак, не враховуючи при цьому загального впливу всього набору ознак на вихідний параметр. Альтернативою їм є методи еволюційного пошуку [4, 5], які на кожній ітерації працюють із множиною потенційних розв'язків одночасно, що дозволяє повніше охопити простір пошуку, ніж дозволяють градієнтні методи оптимізації, та отримати розв'язок, близький до оптимального, за відносно короткий час.

На сьогоднішній день найкращі результати показує метод еволюційного пошуку з кластеризацією ознак [6]. Він враховує близькість ознак у просторі об'єктів аналізу та дозволяє відкинути ознаки, що несуть однакову інформацію про досліджуваний об'єкт. Його основним недоліком є неможливість врахування знань експерта в процесі прийняття рішення про врахування чи неврахування тієї чи іншої ознаки в подальшому аналізі даних.

Метою дослідження є розробка методу автоматичного вибору комбінації інформативних ознак, що інтегрував би в собі використання емпіричних спостережень та експертних знань.

Постановка задачі

Нехай задано навчальний набір даних X, що складається із множини зразків. Для кожного зі зразків експертом у цій галузі задано значення вихідного параметра з множини Y:

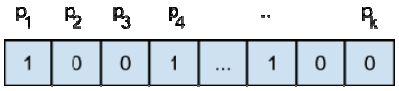
$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ \dots & & & \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}; \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \tag{1}$$

де n — кількість зразків даних набору, $P = \{p_1, \dots, p_k\}$ — універсальна множина ознак, x_{ij}, $i = [1, n]$, $j = [1, k]$ — значення ознаки $p_j \in P = \{p_1, ..., p_k\}$ для зразка $p_j \in P = \{p_1, ..., p_k\}$, y_i — задане експертом значення вихідного параметра для заданого набору значень вхідних параметрів.

Пропонується на основі набору даних (1) визначити таку підмножину ознак $P' \in P$, що найточніше характеризувала би множину X, при цьому повною мірою використовуючи знання експерта про досліджуваний об'єкт. В задачах, що описуються простором ознак високої розмірності, вкрай складно, або й неможливо побудувати систему, здатну коректно працювати, спираючись виключно на експериментальні дані, в той час як експерт може володіти інформацією про приховані закономірності, присутні в них.

Метод еволюційного пошуку інформативних ознак із залученням експерта

Метод еволюційного пошуку інформативних ознак являє собою еволюційний алгоритм, що на кожній ітерації працює з підмножиною потенційних розв'язків. Кожен кандидат на оптимальний розв'язок представлений хромосомою — бітовим рядком з k елементів, де k — загальна кількість усіх можливих ознак, що описують об'єкт. Якщо ознака береться до розгляду в поточному розв'язку-кандидаті, то відповідний біт хромосоми встановлюється рівним 1 (рис.).



Структура хромосоми

Над хромосомами кожного покоління проводяться процедури схрещування та мутації, особливості яких можуть варіюватися залежно від характеру прикладної задачі.

Можливість переходу хромосоми в наступне покоління оцінюється за допомогою критерію оптимізації T, що являє собою суму квадратів відхилень реального значення вихідного параметра (його надає експерт у предметній галузі в форматі «об'єкт належить/не належить до заданого кластера») від значення, обчисленого за допомогою моделі, синтезованої на основі вибраної комбінації ознак.

$$T(C_e, C) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(\mu_i^j - \mu_{e_i}^j \right)^2 \to \min, \tag{2}$$

де $\mu_{ie} \in \{0, 1\}$ — належність об'єкта до кластера, визначена експертом, $\mu_i \in [0, 1]$ — належність об'єкта до кластера за виходом синтезованої моделі.

Критерій (2) оцінює різницю між двома розбиттями того самого набору вхідних даних на клас-

тери [7, 8]. Одне з розбиттів — C_{e_i}, запропоноване експертом для набору даних (1), інше — C_i, отримане за допомогою моделі, синтезованої на основі набору ознак, представленого цим розв'язком. Елементами кластерів виступають зразки x_1, ... x_n; відстань між двома зразками x_1, та x_2 визначається за Евклідовою метрикою:

$$d_E(x_1, x_2) = \sqrt{\sum_{i=1}^{k} \left(x_{1_i} - x_{2_i} \right)^2}, \tag{3}$$

де k — кількість ознак.

Метод кластеризації, використаний в запропонованій реалізації — метод нечітких c-means [9].

По завершенні роботи еволюційного алгоритму виконується процедура оптимізації вибраних ознак за рахунок відкидання надлишкових, тобто таких, що несуть ідентичну інформацію про об'єкт. Для цього проводиться кластеризація вибраних ознак у просторі, осями якого виступають зразки x_1, ... x_n. Елементами кластерів у цьому випадку виступають ознаки. Відстань між двома ознаками p_1, та p_2 також визначається за Евклідовою метрикою:

$$d_E(p_1, p_2) = \sqrt{\sum_{j=1}^{n} \left(x_{j_1} - x_{j_2} \right)^2}, \tag{4}$$

де n — кількість зразків.

У складі хромосоми залишаються лише ознаки, що мають високі ступені належності до кластерів $\mu_{ij} \geq \eta$.

Виконання описаної вище процедури відкидання ознак наприкінці роботи методу має суттєву перевагу. Перед прийняттям остаточного рішення про відкидання деякої множини ознак можна надати її експерту для підтвердження. Експерт може підтвердити виключення ознак, або ж не дозволити вилучення тих, що, на його думку, все ж повинні бути враховані в процесі подальшого аналізу даних. Перелік ознак, що пропонуються експерту на верифікацію, формується на основі ступеня належності ознак до кластеру. До нього потрапляють усі ознаки, що належать до кластеру зі ступенем належності $\mu_{ij} \geq \eta$, де за замовчуванням $\eta = 0{,}7$, але може конфігуруватися користувачем системи (експертом).

Запропонований метод пропонується виконувати як таку послідовність кроків:

1. Випадковим чином згенерувати m хромосом вигляду, як на рис.

2. Для кожної з m хромосом побудувати розбиття зразків із множини X (1) на кластери C_i, враховуючи лише ознаки з одиничним значенням у відповідному біті.

Оцінити значення фітнес-функції (2), маючи розбиття C_i, побудоване на основі поточної хромосоми, та C_{e_i}, задане експертом для набору даних (1).

3. Серед m хромосом у поколінні вибрати $m_1 \leq m$ з імовірністю p_1 та провести над ними операцію одноточкового схрещування. Оцінити значення фітнес-функції для кожного з $\lfloor m_1/2 \rfloor \cdot 2$ нащадків.

4. Серед m хромосом у поколінні вибрати $m_2 \leq m$ з імовірністю p_2 та провести над ними операцію двоточкового схрещування. Оцінити значення фітнес-функції для кожного з $\lfloor m_2/2 \rfloor \cdot 2$ нащадків.

5. Серед m хромосом у поколінні вибрати $m_3 \leq m$ з імовірністю p_3 та провести над ними операцію мутації шляхом інверсії одного з генів хромосоми [10]. Оцінити значення фітнес-функції для кожного з m_3 мутантів.

6. Серед початкових хромосом із популяції, нащадків обох операцій схрещування та мутантів відібрати m особин із найменшими значеннями фітнес-функції.

7. Повторювати кроки 2—6 до збігання популяції або завершення часу, відпущеного на еволюцію. Збіганням вважатимемо ситуацію, коли повторення кроків 2—6 не веде до зменшення значення фітнес-функції протягом l ітерацій.

У цій реалізації пропонується брати $m = 20$; $p_1 = 0{,}8$; $p_2 = 0{,}5$; $p_3 = 0{,}25$; $l = 5$.

8. Вибрати з останнього покоління хромосому з мінімальним значенням фітнес-функції. Для кожної пари ознак, що мають одиничні значення у відповідних бітах хромосоми, обчислити Евклідові відстані (4) в просторі зразків.

9. На основі отриманих Евклідових відстаней провести кластеризацію ознак методом c-means.

За отриманим розбиттям вибрати ознаки, найближчі до центрів кластерів: $\mu_{ij} \geq \eta$.

10. Надати експерту для підтвердження перелік ознак, які пропонується виключити з розгляду. Сформувати остаточний перелік інформативних ознак.

Така реалізація методу еволюційного пошуку інформативних ознак дозволяє повною мірою врахувати знання експерта, що мають особливо високу цінність у задачах високої розмірності. Разом з тим, як буде показано далі, вона спричиняє порівняно значну похибку. Далі пропонується прийом, що дозволить знизити вплив похибки роботи методу та підвищити стабільність його роботи.

Метод еволюційного пошуку інформативних ознак (із паралельною еволюцією)

Для підвищення стабільності роботи методу пропонується використати паралельну еволюцію двох популяцій. По завершенні еволюційної процедури виконується об'єднання обох підмножин інформативних ознак, з якого, в свою чергу, виключаються надлишкові ознаки за вищепоказаним алгоритмом.

В загальному випадку ця остання процедура повинна виконуватись приблизно вдвічі довше за попередню. Проте, з технічної точки зору переважна більшість сучасних комп'ютерних систем мають мультипроцесорну архітектуру, а остання із запропонованих реалізацій пропонує ширші можливості для розпаралелювання процесу виконання, тому втрати швидкодії на практиці незначні. В першому методі розпаралелювання можливе лише на рівні підрахунку значень фітнес-функції для різних особин у популяції; в другому ж, природно, — на рівні самих популяцій.

Демонстраційний приклад та порівняння різних підходів

Для демонстрації можливостей обох реалізацій методу, описаного раніше в цій роботі, було використано простір ознак, що складається з 46 параметрів, які характеризують артезіанську свердловину: клас за хімічним складом; рівень мінералізації; відстань до найближчого міста; відстань до залізничних колій; глибина свердловини; середньорічна кількість опадів; потужність водоносного горизонту; статичний рівень води; коефіцієнт фільтрації; загальна жорсткість; рівень pH; річна амплітуда коливань рівня води; макрокомпонентний склад: Na^+, K^+, Ca^{+2}, Mg^{+2}, Cl^-, SO_4, HCO_3^-; мікрокомпоненти: Al, Ba, … — всього 21; пестициди (ДДТ, метафос — всього 6).

Пропонується з усього набору ознак виявити підмножину з $k \geq 10$ таких, щоб мінімізувати критерій (3), виходячи з навчального набору даних з 20 векторів. Приклад вхідних векторів показано в табл. 1.

Таблиця 1

Приклад вхідних векторів навчального набору даних

Клас за хімічним складом	Рівень мінералізації	Відстань до найближчого міста	Відстань до залізничних колій	Глибина свердловини	…	Якість води
гідрокарбонатна магнієво-кальцієва	0,4—0,6 г/дм3	20 км	2 км	100 м	…	добра
гідрокарбонатна кальцієво-магнієва	0,4—0,7 г/дм3	1 км	1 км	120 м	…	задовільна
. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	...	. . . . .

По завершенні виконання процедури оптимізації розв'язок P' з мінімальним значенням критерію T = 10,33 включав такі ознаки: рівень мінералізації; середньорічна кількість опадів; загальна жорсткість; рівень pH; макрокомпонентний склад: Na^+, Cl^-, HCO_3^-; мікрокомпоненти: Pb, Mn, Rn, Ag.

Порівняльну характеристику двох вищерозглянутих реалізацій методу наведено в табл. 2. Всі значення є середніми за 10 запусків на кожен із методів. Порівняння здійснювалось за такими критеріями:

t — час, витрачений на пошук оптимальної комбінації інформативних ознак;

k_f — кількість підрахованих значень фітнес-функції;

k — кількість відібраних ознак;

$\varepsilon = \dfrac{d\left(P_n, P_c\right)}{k}$ — досягнутий рівень похибки, де $d(P_n, P_c)$ — відстань Хемінга між розв'язком, об-

раним із отриманих за синтезованими моделями та розв'язком, запропонованим експертом.

Таблиця 2

Порівняльна характеристика методів

Назва методу	Критерій			
	t, с	k_f	k	ε
Метод еволюційного пошуку інформативних ознак із залученням експерта	92,02	99,25	11,2	0,1739
Метод еволюційного пошуку інформативних ознак (із паралельною еволюцією)	96,10	202,54	11,4	0,0652

Із табл. 2 випливає, що друга реалізація дає можливість підвищити якість отриманих результатів до задовільного рівня за незначного зниження швидкості роботи. Значення похибки можна вважати незначним, оскільки вона еквівалентна в середньому одній неправильно визначеній ознаці з 46 заданих.

Висновки

1. Запропоновано метод автоматичного вибору комбінації інформативних ознак, що дозволяє:

— виявити зв'язки між вхідними та вихідним параметрами моделі;

— скоротити простір ознак за рахунок таких, що не чинять впливу на значення вихідного параметра, або впливом яких можна знехтувати;

— з надходженням нових даних система може продовжувати навчатися, і результуючий набір інформативних ознак може змінитися.

2. Запропонований метод інтегрує емпіричні спостереження (навчальний набір даних) та експертні знання (значення порогів, контроль експерта за відкиданням ознак, позначених системою як надлишкові).

3. Дозволяє уникнути помилкового виключення важливих ознак в результаті неповноти даних у навчальному наборі або невдалого вибору порогу ступенів належності.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Субботин С. А. Неитеративный синтез нейро-нечетких диагностических экспертных систем / С. А. Субботин // Штучний інтелект. — 2009. — № 4. — С. 380—386.

2. 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. : ил.

3. Прикладная статистика: Классификация и снижение размерности : справ. изд. / [С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин]. — М. : Финансы и статистика, 1989. — 607 с.

4. Haupt R. Practical Genetic Algorithms / R. Haupt, S. Haupt. — New Jersey : John Wiley & Sons, 2004. — 261 p.

5. Subbotin S. Entropy Based Evolutionary Search for Feature Selection / S. Subbotin, A. Oleynik // The experience of designing and application of CAD systems in Microelectronics : Proceedings of the IX International Conference CADSM-2007 (20—24 February 2007). — Lviv : Publishing house of Lviv Polytechnic, 2007. — P. 442—443.

6. 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.

7. Oliveira J. V. Advances in Fuzzy Clustering and Its Applications / J. V. Oliveira, W. Pedrycz. — John Wiley & Sons Ltd., 2007. — 435 pp.

8. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл. — М. : Статистика, 1977. — 128 с.

9. Bezdeck J. C. FCM : Fuzzy C-Means Algorithm / J. C. Bezdeck, R. Ehrlich, W. Full // Computers and Geoscience. — 1984. — № 10 (2—3). — P. 191—203.

10. Ротштейн А. П. Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети / А. П. Ротштейн. — Винница : Универсум-Винница, 1999. — 320 с.

Рекомендована кафедрою захисту інформації ВНТУ

Стаття надійшла до редакції 24.02.2015

Кондратенко Наталія Романівна — канд. техн. наук, доцент, професор кафедри захисту інформації;
Снігур Ольга Олексіївна — аспірантка кафедри захисту інформації, e-mail: olha.snihur@gmail.com.
Вінницький національний технічний університет, Вінниця

N. R. Kondratenko[1]
O. O. Snihur[1]

Expert-involving evolutionary search of informative features for evaluation of quality of artesian water

[1]Vinnytsia National Technical University

The problem of detecting an informative feature set, that describes the condition of an artesian well, is being solved. The method of automatic selection of features combination is introduced. The method integrates the usage of empirical observations and expert knowledge and enables reduction of the feature space by excluding features, that have no effect on the output parameter value, or the effect of which can be neglected.

Keywords: decision making support system, expert system, Data Mining, clustering, informative features detection, hydrogeology, evolutionary search.

Kondratenko Natalia R. **—** Cand. Sc. (Eng.), Assistant Professor, Professor of the Chair of Information Security;
Snihur Olha O. — Post-Graduate Student of the Chair of Information Security, e-mail: olha.snihur@gmail.com

Н. Р. Кондратенко[1]
О. О. Снигур[1]

Эволюционный поиск информативных признаков с привлечением эксперта в задаче оценки качества артезианской воды

[1]Винницкий национальный технический университет

Решена задача выделения набора информативных признаков, описывающих состояние артезианской скважины. Предложен метод автоматического выбора комбинации признаков, интегрирующий эмпирические наблюдения и экспертные знания и позволяющий сократить пространство признаков за счет не имеющих влияния на значение выходного параметра, или влиянием которых можно пренебречь.

Ключевые слова: система поддержки принятия решений, экспертная система, интеллектуальный анализ данных, кластеризация, выделение информативных признаков, гидрогеология, эволюционный поиск.

Кондратенко Наталия Романовна — канд. техн. наук, доцент, профессор кафедры защиты информации;
Снигур Ольга Алексеевна — аспирант кафедры защиты информации, e-mail: olha.snihur@gmail.com