УДК 004.89+336.713

Т. О. САВЧУК, С. І. ПЕТРИШИН

Вінницький національний технічний університет, Вінниця

ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ МОДЕЛЮВАННЯ ПРОЦЕСУ КЛАСТЕРНОГО АНАЛІЗУ НАДЗВИЧАЙНИХ СИТУАЦІЙ НА ЗАЛІЗНИЧНОМУ ТРАНСПОРТІ

Анотація. У статті проведено оцінювання результатів моделювання процесу кластерного аналізу надзвичайних ситуацій на залізничному транспорті за допомогою програмного продукту , що базується на модифікованому алгоритмі k-means та програмного засобу DEDUCTOR.

Ключові слова: моделювання, Data Mining, надзвичайна ситуація, залізничний транспорт, кластеризація, кластер.

Аннотация. В статье проведено оценивание результатов моделирования процесса кластерного анализа чрезвычайных ситуаций на железнодорожном транспорте с помощью программного продукта, основанном на модифицированном алгоритме k-means и программного средства DEDUCTOR.

Ключевые слова: моделирование, Data Mining, чрезвычайная ситуация, железнодорожный транспорт, кластеризация, кластер.

Abstract. The paper carried the results of evaluation of process modeling of cluster analysis of emergencies on the railways with the help of a software product based on a modified k-means algorithm and software tools DEDUCTOR.

Key words: Modeling, Data Mining, emergencies, rail transport, clustering, cluster.

Вступ

Характеристики оцінювання результатів моделювання кластерного аналізу надзвичайних ситуацій як складного процесу [1, 2] визначаються як властивостями його складових, так і характером взаємодії між ними. Слід врахувати такі особливості процесу оцінювання:

- його стан описується потужним вектором динамічних змінних;
- виявляє якісні зміни поведінки його складових;
- включає нелінійні взаємодії основних його складових і обернені зв'язки між ними, які також містять нелінійності.

Серед основних задач, що є актуальними в процесі оцінювання результатів кластеризації надзвичайних ситуацій на залізничному транспорті, є розробка його моделі шляхом відтворення зв'язків і відношень між основними його складовими [2, 3].

Для дослідження процесу оцінювання результатів кластерного аналізу надзвичайних ситуацій на залізничному транспорті з використанням фізичного моделювання, як одного з найбільш поширених на практиці підходів [2], в якості фізичної моделі може виступати:

- процес фізичної природи, що описується аналогічним математичним апаратом;
- процес аналогічної фізичної природи, але в другій області параметрів (масштабна модель).

Оскільки підбір процесу фізичної природи аналогічного процесу оцінювання результатів кластерного аналізу і відповідного математичного апарату, є складним, можна зробити висновок про недоцільність застосування даного виду моделювання для розв'язання поставленої задачі [2].

Іншим підходом, також достатньо популярним в практиці дослідження, є математичне моделювання, яке дозволяє оцінювати результати кластерного аналізу надзвичайних ситуацій на залізничному транспорті, враховуючи його якісні та кількісні характеристики, що є доцільним для проведення оцінювання результатів функціонування програмних засобів, що базуються на відповідних алгоритмах.

Отже, оцінювання результатів кластерізації за допомогою засобів, що базуються на класичному та модифікованому алгоритмах k-means для аналізу надзвичайних ситуацій на залізничному транспорті, доцільно проводити за допомогою математичного моделювання.

Постановка задачі

Провести оцінювання результатів моделювання процесу кластерного аналізу надзвичайних ситуацій за допомогою програмного засобу DEDUCTOR для аналізу таких ситуацій, який базується на класичному алгоритмі k-means, та розробленого програмного засобу, що базується на модифікованому алгоритмі k-means [4] (в подальшому – Program analysis of emergency situations (PAES)) з метою визначення якості кластеризації при аналізі надзвичайних ситуацій на залізничному транспорті.

Формалізація процесу оцінювання результатів кластеризації надзвичайних ситуацій на залізничному транспорті

Оцінювання результатів кластеризації надзвичайних ситуацій на залізничному транспорті можна здійснити шляхом використання критеріальних величин.

Існує потужна множина критеріїв, що можуть бути використані для аналізу надзвичайних ситуацій на залізничному транспорті, серед них було обрано частку загального розкиду, точково-бісеріальний метод кореляції та узагальнену дисперсію в класах, яких достатньо для оцінки якості результатів розбиття [5].

Нехай множина надзвичайних ситуацій на залізничному транспорті розбита на *k* кластерів $G_i \ (i = \overline{1,k})$.

Для визначення частки загального розкиду надзвичайних ситуацій між кластерами *T* необхідно ввести такі три характеристики ступеню розсіювання надзвичайних ситуацій [1, 5, 6]:

- загальне розсіювання *S*, що визначається як

$$S = \sum_{i=1}^{n} a^2(Y_i, \overline{Y}), \tag{1}$$

де Y_i - вектор даних про *i*-ту надзвичайну ситуацію;

$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ – загальний центр ваги;

n – кількість надзвичайних ситуацій, що аналізуються;

$a^2(Y_i, \overline{Y})$ – квадрат відстані між *i*-ю надзвичайною ситуацією та загальним центром ваги;

- міжкластерний розкид *B*

$$B = \sum_{z=1}^{k} n_z a^2(\overline{Y_z}, \overline{Y}), \tag{2}$$

де $\overline{Y_z} = \frac{1}{n_z}\sum_{Y_i \in G_z} Y_i$ – центр ваги *z*-го кластера надзвичайних ситуацій на залізничному транспорті;

n_z – кількість надзвичайних ситуацій на залізничному транспорті в кластері G_z;

$a^2(\overline{Y_z}, \overline{Y})$ – квадрат відстані між центром ваги *z*-го кластера та загальним центром ваги;

- розкид всередині кластерів надзвичайних ситуацій на залізничному транспорті

$$W = \sum_{z=1}^{k} W_z, \tag{3}$$

де $W_z = \sum_{Y_i \in G_z} a^2(Y_i, \overline{Y})$.

Оскільки кластерний аналіз надзвичайних ситуацій базується на модифікованому алгоритмі k-means [4], то

$$S = W + B. \tag{4}$$

Тоді, частка загального розкиду надзвичайних ситуацій на залізничному транспорті *T* може бути визначена як [4, 5]:

$$T = 1 - \frac{W}{S} \tag{5}$$

Частка загального розкиду надзвичайних ситуацій *T* є нормованою величиною $(0 \le T \le 1)$. При цьому, наближення значення до 0 свідчить про нижчу якість розбиття надзвичайних ситуацій на залізничному транспорті на таксони.

Точково-бісеріальний коефіцієнт кореляції R_b між надзвичайними ситуаціями, що аналізуються, визначається таким чином. Кожній парі надзвичайних ситуацій на залізничному транспорті Y_i та Y_j ставиться у відповідність дві величини – відстань між ними та індекс еквівалентності δ_{ij} [4, 5]

$$\delta_{ij} = \begin{cases} 1, \text{якщо } Y_i \text{ та } Y_j \text{ належать одному кластеру;} \\ 0, \text{в протилежному випадку.} \end{cases} \quad (6)$$

Коефіцієнт R_b підраховується як коефіцієнт кореляції між a_{ij} та бінарною величиною δ_{ij} по всіх парах надзвичайних ситуацій, які аналізуються, що дає [5, 6]

$$R_b = \frac{(\overline{a_b} - \overline{a_w})\sqrt{\frac{f_w f_b}{n_a^2}}}{s_a}, \quad (7)$$

де $\overline{a_b}$ – середня відстань між надзвичайними ситуаціями із різних кластерів;

$\overline{a_w}$ – середня відстань між надзвичайними ситуаціями із одного кластера;

f_w - кількість відстаней між надзвичайними ситуаціями, що потрапили в однин кластер;

f_b - кількість відстаней між надзвичайними ситуаціями із різних кластерів;

n_a - загальна кількість відстаней;

s_a - стандартне відхилення відстаней.

Узагальнена дисперсія в класах надзвичайних ситуацій на залізничному транспорті H є однією з характеристик ступеню розсіювання надзвичайних ситуацій на залізничному транспорті, що належать одному класу, і обраховується за формулою [5, 6]:

$$H = \det(\sum_{l=1}^{k} n_l W_l), \quad (8)$$

де $\det(\sum_{l=1}^{k} n_l W_l)$ - визначник матриці.

Елементи $w_{qm}(l)$ вибіркової коваріаційної матриці W_l можуть бути визначениі як

$$w_{qp}(l) = \frac{1}{n_l} \sum_{Y_i \in G_l} (y_i^{(q)} - \overline{y}^{(q)}(l))(y_i^{(p)} - \overline{y}^{(p)}(l)), \; q, p = 1, 2, ..., m, \quad (9)$$

де $y_i^{(p)}$ – p-та характеристика надзвичайної ситуації на залізничному транспорті Y_i;

$\overline{y}^{(p)}(l)$ – середнє значення p-ї компоненти, підраховане за надзвичайними ситуаціями l-го класу.

Відносний показник якості розбиття множини надзвичайних ситуацій на таксони може бути визначеним як

$$K = \frac{\overline{T'} + \overline{R_b'} + \overline{H'}}{3} \quad (10)$$

де T', R_b' і H' - відносні значення показників T, R_b і H.

Таким чином, процес оцінювання результатів кластеризації надзвичайних ситуацій на залізничному транспорті буде включати такі етапи (рисунок 1):

1) моделювання процесу кластерного аналізу за допомогою за допомогою програмного засобу DEDUCTOR та за допомогою PAES;
2) обрахування частки загального розкиду T надзвичайних ситуацій між кластерами;
3) обрахування точково-бісеріального коефіцієнта кореляції R_b між надзвичайними ситуаціями, що аналізуються;
4) обрахування узагальненої дисперсії в класах надзвичайних ситуацій на залізничному транспорті H;
5) обрахування відносного показника якості розбиття множини надзвичайних ситуацій на таксони.

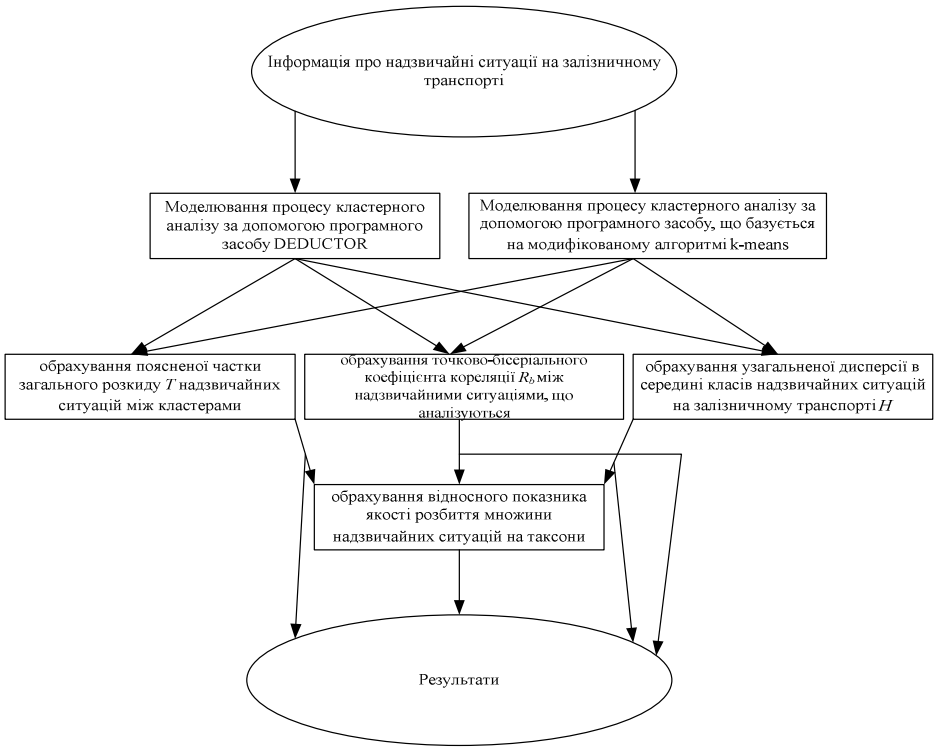

Рисунок 1 - Етапи процесу оцінювання результатів кластеризації надзвичайних ситуацій на залізничному транспорті

Оцінювання результатів процесу кластеризації надзвичайних ситуацій на залізничному транспорті

При використанні програмного засобу DEDUCTOR для аналізу надзвичайних ситуацій та PAES, було проведено 3 експерименти:

- в першому – було обрано 5 значущих параметрів та характеристик надзвичайних ситуацій, що в повній мірі відображають стан таких ситуацій;
- в другому – окрім цих п’яти було додано ще одну – менш значущу;
- в третьому – окрім значущих було додано ще 5 не важливих характеристик та параметрів при ідентифікації надзвичайних ситуацій.

Частку загального розкиду T (таблиця 1) визначений за результатами моделювання процесів кластерного аналізу надзвичайних ситуацій показав, що якість розбиття, яке виконане PAES є вищою, оскільки $T_{\text{PAES}} > T_{Deductor}$ (рисунок 2).

Точково-бісеріальний коефіцієнт кореляції R_b (таблиця 1) визначений за результатами моделювання процесів кластерного аналізу надзвичайних ситуацій показав, що якість розбиття, яке виконане програмним засобом DEDUCTOR для аналізу надзвичайних ситуацій є вищою, оскільки $R_{bDeductor} > R_{b\text{PAES}}$, проте різниця між отриманими значеннями є незначною та за певних умов можна вважати, що, $R_{bDeductor} \geq R_{bPAES}$, або $R_{bDeductor} = R_{b\text{PAES}}$, а отже, можна зробити висновок про те, що якість розбиття в обох випадках (за даним критерієм) є однаковою (рисунок 2).

Таблиця 1 – Оцінювання результатів кластерного аналізу надзвичайних ситуацій на залізничному транспорті

Засіб	Алгоритм кластеризації	Кількість кластерів			Частка загального розкиду, T			Точково-бісеріальний коефіцієнт кореляції, R_b			Узагальнена дисперсія в класах, H			Відносний показник якості розбиття
Номер експерименту		1	2	3	1	2	3	1	2	3	1	2	3	
Програмний засіб DEDUCTOR для аналізу надзвичайних ситуацій	k-means	3	3	4	0,8005	0,8018	0,8399	0,586	0,516	0,469	0,721	0,744	0,782	75,4%
PAES	Модифікований k-means	3	3	3	0,8199	0,8233	0,8475	0,563	0,512	0,482	0,672	0,689	0,715	78,31%

Узагальнена дисперсія в класах H (таблиця 1) визначена за результатами моделювання процесів кластерного аналізу надзвичайних ситуацій показала, що якість розбиття, яке виконане програмним засобом DEDUCTOR для аналізу надзвичайних ситуацій є нижчою, оскільки $H_{Deductor} > H_{\text{PAES}}$, що свідчить про більші відхилення від центра кластерів надзвичайних ситуацій при розбитті програмним засобом DEDUCTOR для аналізу надзвичайних ситуацій (рисунок 2).

Відносний показник якості розбиття множини надзвичайних ситуацій на таксони (таблиця 1) визначений за результатами моделювання процесів кластерного аналізу надзвичайних ситуацій показав, що якість розбиття, яке виконане програмним засобом DEDUCTOR для аналізу надзвичайних ситуацій є нижчою на 2,91%, що свідчить про меншу якість розбиття множини надзвичайних ситуацій на кластери програмним засобом DEDUCTOR для аналізу надзвичайних ситуацій (рисунок 2).

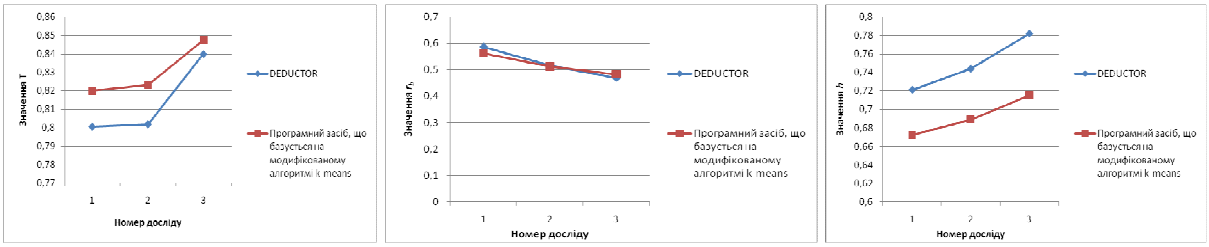

Рисунок 2 - Порівняння результатів моделювання процесу кластерного аналізу надзвичайних ситуацій на залізничному транспорті

Проведене дослідження показало, що удосконалення класичного алгоритму k-means шляхом введення нових параметрів до цільової функції привело до підвищення якості кластеризації надзвичайних ситуацій на залізничному транспорті на 2,9%, а отже, модифікований алгоритм доцільно застосовувати для аналізу надзвичайних ситуацій на залізничному транспорті.

Висновки

Таким чином, за результатами моделювання процесу оцінювання результатів кластерного аналізу надзвичайних ситуацій на залізничному транспорті можна зробити висновок, що модифікований алгоритм k-means має такі переваги по відношенню до класичного алгоритму k-means, що покладений в основу програмного засобу DEDUCTOR, як

1) якість розбиття за часткою загального розкиду T є вищою;
2) якість розбиття за узагальненою дисперсією в класах H є вищою;
3) у випадку точково-бісеріального коефіцієнта кореляції R_b – якість розбиття практично однакова;
4) за відносним показником якості розбиття – модифікований алгоритм k-means має перевагу 2,9% при кластеризації надзвичайних ситуацій на залізничному транспорті в порівнянні із звичайним алгоритмом k-means.

Список літератури

1. Савчук Т.О., Петришин С.І. Визначення евклідової відстані між надзвичайними ситуаціями на залізничному транспорті під час кластерного аналізу//Наукові праці Вінницького національного технічного університету. – Серія «Інформаційні технології та комп'ютерна техніка». – 2010. – Випуск №3(2010). - http://www.nbuv.gov.ua/e-journals/vntu/2010_3/2010

2. В.М. Томашевский Моделювання систем – К. Видавнича група BHV, 2005 – 352 с.
3. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004 – 336с.
4. Савчук Т.О., Петришин С.І. Розробка модифікованого алгоритму K-MEANS для аналізу надзвичайних ситуацій на залізничному транспорті// Матеріали конференції, Обчислювальний інтелект (результати, проблеми, перспективи): Матеріали 1-ї Міжнародної науково-технічної конференції (10-13 травня 2011 р.), - Черкаси, 2011, - С. 236-237.
5. Айвазян С.А., Бухштабер В.М., Енюков И.С. Прикладная статистика: Классификация и снижение размерности. – М.:Финансі и статистика, 1989. – 607 с.
6. Мандель И.Д. Кластерный анализ. – М.:Финансы и статистика, 1988. – 176с.

Стаття надійшла: 28.03.12.

Відомості про авторів

Савчук Тамара Олександрівна – кандидат технічних наук, доцент, професор кафедри Комп’ютерних наук, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, тел.0664124037, savchtam@rambler.ru

Петришин Сергій Іванович – аспірант кафедри Комп’ютерних наук Вінницького національного технічного університету, Хмельницьке шосе, 95, м Вінниця, 21021.