УДК 004.89+336.713

# Порівняльний аналіз використання методів кластеризації для ідентифікації надзвичайних ситуацій на залізничному транспорті

Т.О. Савчук, С.І. Петришин
Вінницький національний технічний університет
savchtam@rambler.ru

***Abstract***

***T.A. Savchuk, S.I. Petrishyn Comparative analysis using clustering methods to identify emergency situations on railway.*** *An approach to generate rules, which will be clustering of emergencies on the railway to its identity. Advantages of the algorithm: high performance, clustering can in real time. The disadvantage of the algorithm: the possibility of false solutions in the analysis of fundamentally new emergencies.*

Техніка кластеризації застосовується в різних областях [1]. В аналізі надзвичайних ситуацій на залізничному транспорті достовірна їх кластеризація може стати вирішальним кроком для зменшення наслідків або взагалі запобігання їх виникнення. Якщо необхідно розподілити велику кількість інформації за придатними для подальшої обробки групами, що є актуальним для оперативного аналізу надзвичайних ситуацій на залізничному транспорті з метою їх ідентифікації, кластерний аналіз виявляється досить ефективним [2].

Для вирішення задачі аналізу надзвичайної ситуації на залізничному транспорті з метою її ідентифікації можна використовувати класифікацію, але об’єкти при цьому розподіляються по певних класах за вже відомими, наперед визначеними ознаками, що було б ідеальним при оперативному аналізі надзвичайних ситуацій, але це неможливо, оскільки параметри аналізуємої ситуації можуть бути не повними, або введені наперед не визначені параметри, що є передумовою для недостовірної класифікації. В процесі кластерного аналізу класи, до яких відносяться в подальшому об’єкти (надзвичайні ситуації), формуються на основі їх властивостей (такий аналіз називають навчанням без вчителя [3]), що при неможливості передбачити всі можливі класи таких ситуацій на залізничному транспорті є актуальним для їх ідентифікації, та основною перевагою кластеризації в порівнянні з класифікацією при розв’язанні даної задачі.

## *Визначення доцільності використання кластерного підходу для аналізу надзвичайних ситуацій на залізничному транспорті*

Ключовим поняттям кластерного аналізу надзвичайних ситуацій на залізничному транспорті є подібність конкретних надзвичайних ситуацій, яка математично виражається за допомогою міри їх схожості [4]. На основі значення цієї міри формується висновок про близькість надзвичайних ситуацій і приймається рішення про їх належність одному кластеру. Опис даних на основі об’єктів та ознак в ході кластерного аналізу зазвичай втрачається, його замінює матриця схожості надзвичайних ситуацій.

При цьому, в самих кластерах загальний ознаковий опис складових їх об'єктів явно не виражений, а це призводить до появи некоректних класів надзвичайних ситуацій (наприклад, горючі та негорючі речовини виявляються в одному кластері). Проте в термінах ознакового опису може з'ясуватися, що ці речовини наприклад є рідинами, мають однаковий колір та густину та ін. Така помилка може призвести до серйозних негативних наслідків – неможливість інтерпретації результатів експериментів загрожує прийняттю недостовірних рішень при виконанні робіт по ліквідації наслідків надзвичайної ситуації. Означені недоліки методів кластеризації можуть відігравати важливу роль як для навколишнього середовища, так і для населення. До цього призводить той факт, що існує множина типових надзвичайних ситуацій, в якій кожна аналізуєма схожа в чомусь з сусідніми, але схожість ця не транзитивна, а тому загальний ознаковий опис

даних надзвичайних ситуацій при обчисленні схожості на кожному кроці не враховується. В результаті може бути некоректно ідентифіковано надзвичайну ситуацію, що призведе до негативних наслідків, та прийняття хибних рішень при формуванні ліквідаційних підрозділів або координації їх дій

Отже, кластерний аналіз може бути використаним при розв'язанні таких основних типів задач аналізуємої предметної області [5]:

1)розробки типології або класифікації надзвичайних ситуацій на залізничному транспорті;

2) дослідження різних корисних концептуальних схем групування надзвичайних ситуацій на залізничному транспорті;

3) створення гіпотез на основі дослідження даних про такі ситуації;

4) перевірка гіпотез або дослідження для визначення, чи справді групи, виділені тим чи іншим способом, присутні в наявних даних про надзвичайну ситуацію.

Аналіз даних про швидкоплинні надзвичайні ситуації з метою їх ідентифікації включає такі основні кроки [6]:

1) відбір вибірки надзвичайних ситуацій для кластеризації;

2) визначення множини ознак, за якими будуть оцінюватись надзвичайні ситуації на залізничному транспорті у вибірці;

3) обрахування значень мір схожості між надзвичайними ситуаціями;

4) застосування методу кластерного аналізу для створення груп подібних надзвичайних ситуацій на залізничному транспорті;

5) перевірка достовірності результатів кластерного аналізу.

Однак слід відзначити такі особливості використання кластерного підходу для аналізу надзвичайних ситуацій на залізничному транспорті:

– Значна кількість методів кластерного аналізу – це прості процедури, які не мають достатнього статистичного обгрунтування. Це означає, що більшість методів такого аналізу є евристичними, тобто підкріплені лише власним досвідом розробників. Отже, для визначення найбільш доцільного методу кластеризації для аналізу надзвичайних ситуацій потрібно проаналізувати методи, які будуть застосовуватись, чи вони не є евристичними і дійсно дієвими.

– Методи кластерного аналізу розроблялись для багатьох наукових дисциплін (наприклад для біології, психології та ін) і, як наслідок, містять в собі їх особливості, що неважливі при аналізі надзвичайних ситуації на залізничному транспорті. Отже, для аналізу таких ситуацій доцільно використовувати специфічні для аналізуємої предметної області підходи кластерного аналізу.

– Різні кластерні методи породжують різні варіанти кластерів для одних і тих самих наборів даних. Даний недолік може негативно вплинути на аналіз надзвичайних ситуацій, що призведе до негативних наслідків для населення та навколишнього середовища. Отже, слід провести ряд тестів, які підтвердять доцільність використання обраного методу для аналізу надзвичайних ситуацій.

### *Постановка задачі*

Задачу кластеризації можна визначити в такий спосіб.

В загальному випадку задача кластеризації надзвичайних ситуацій зводиться до того, щоб всю сукупність надзвичайних ситуацій $S=\{S_i\}(\overline{i=1,n})$, статично представлену у вигляді матриці $Y$ або $\psi$, розбити на порівняно невелике число (наперед відоме або ні) однорідних, в певному сенсі, груп, класів або кластерів.

$Y$ – матриця, в якій кожен стовпчик $\{y_i^{(1)},...,y_i^{(j)},...y_i^{(m)}\}$ описує певну надзвичайну ситуацію, тобто $y_i^j$ – певна характеристика окремої надзвичайної ситуації. $\Psi$ – матриця попарної схожості надзвичайних ситуацій на залізничному транспорті. $\Psi$ – квадратна матриця вигляду

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} & \dots & \psi_{1n} \\ \psi_{21} & \psi_{22} & \dots & \psi_{2n} \\ \dots & \dots & \dots & \dots \\ \psi_{n1} & \psi_{n2} & \dots & \psi_{nn} \end{pmatrix},$$

де $\psi_{ij}$ – характеристика схожості (близькості або віддаленості) між об'єктами $S_i$ та $S_j$.

В залежності від способу задання об'єктів можна по-різному інтерпретувати надзвичайні ситуації в вигляді точок: в випадку коли така ситуація задана матрицею $Y$ – ці точки є безпосереднім геометричним зображенням показників $Y_1$, $Y_2$, …, $Y_n$ в $m$-вимірному просторі *Space(Y)*; у випадку з матрицею $\Psi$ – неможливо визначити безпосередньо координати таких точок, проте, на основі значень схожості

(відстаней) між об'єктами виконується даний аналіз.

Для аналізу надзвичайних ситуацій на залізничному транспорті доцільно використати матрицю *Y*, оскільки аналіз має виконуватись в реальному часі, а розрахунок значень $\psi_{ij}$ – окрема задача, для розв'язання якої потрібен час.

Перед розв'язанням задачі кластеризації для аналізу надзвичайних ситуацій на залізничному транспорті потрібно також чітко визначитись з її типом:

– задача розбиття статичного *m*-вимірного діапазону зміни значень аналізуємих ознак на інтервали групування;

– задача визначення природного розшарування вихідних аналізуємих даних на чітко виражені кластери.

Перша задача має розв'язок у будь-якому випадку. При розв'язанні другої - може виникнути ситуація, коли неможливо визначити природного розшарування на кластери, що є неприйнятним для даної предметної області. А це означає, що буде розв'язуватись задача розбиття статичного *m*-вимірного діапазону зміни значень аналізуємих ознак надзвичайних ситуацій на інтервали групування.

Отже, на основі зазначеного можливо сформулювати таку задачу:

Є множина надзвичайних ситуацій на залізничному транспорті $S=\{S_i\}(i=\overline{1,n})$ (дані про еталонні надзвичайні ситуації зберігаються в сховищі даних, тобто всі значення $y_i^j$ є перевіреними і достовірними, а дані про досліджувану надзвичайну ситуацію на залізничному транспорті вводяться ззовні, тобто є ймовірність недостовірних даних), статично представлених у вигляді матриці *Y*. Кожна з надзвичайних ситуацій має *m* характеристик. Потрібно розбити статичний *m*-вимірний діапазон зміни значень аналізуємих ознак надзвичайних ситуацій на інтервали групування. Тобто, множину *S* розбити на *k* ($k<n$) кластерів таким чином, щоб конкретна надзвичайна $Y_i$ ситуація належала одному і тільки одному кластеру, а також щоб надзвичайні ситуації, що належать одному кластеру були максимально подібними а такі ситуації, що належать різним кластерам – максимально різнорідними.

### *Порівняльний аналіз використання методів кластеризації для ідентифікації надзвичайних ситуацій на залізниці*

Всі методи кластерного аналізу поділяються на ієрархічні та неієрархічні [7]. В свою чергу ці групи методи поділяються на підгрупи, кожна з яких є розвиненою, і широко застосовується для різних предметних областей [5].

Перед проведенням кластеризації надзвичайних ситуацій на залізничному транспорті потрібно оцінити різні її методи, та визначитись, якому із них надати перевагу, щоб отриманий результат був найбільш достовірним. Обираючи між ієрархічними і неієрархічну методами, необхідно враховувати такі їх особливості [2]:

– Неієрархічні методи виявляють більш високу стійкість по відношенню до шумів і викидів, некоректного вибору метрики, включенню незначущих змінних в набір, який бере участь в кластеризації, що є актуальним для аналізу надзвичайних ситуацій на залізничному транспорті, оскільки дані будуть вводитись в реальному часі, що збільшує ймовірність виникнення помилок в опрацьовуваних даних. Також до переваг таких методів відноситься те, що за їх допомогою можливо опрацьовувати потужні бази даних, що також є потрібним для опрацювання інформації про надзвичайні ситуації, оскільки при більшій вибірці можливо отримати більш достовірні рішення. Проте основним недоліком неієрархічних методів є те, що на вхід потрібно задавати кількість кластерів або кількість ітерацій, хоча для аналізу надзвичайних ситуацій можливо точно визначити кількість кластерів, оскільки аналізуватись буде лише одна (або декілька) поточна (поточні) надзвичайна ситуація на залізничному транспорті, а всі інші члени вибірки будуть взяті із сховища даних і кількість кластерів буде визначено.

– Якщо немає припущень щодо числа кластерів або кількості ітерацій, потрібно використовувати ієрархічні алгоритми. Однак якщо обробляється потужна база даних, можливий шлях – проведення ряду експериментів з різною кількістю кластерів, наприклад, почати розбиття сукупності даних з двох груп і, поступово збільшуючи їх кількість, порівнювати результати. За рахунок такої зміни результатів досягається досить велика гнучкість кластеризації. Ієрархічні методи, на відміну від неієрархічних, відмовляються від визначення

кількості кластерів, а будують повне дерево вкладених кластерів, що не є особливо актуальним для аналізу надзвичайних ситуацій на залізничному транспорті. Складності ієрархічних методів кластеризації: обмеження обсягу набору даних, оскільки для обробки великої кількості даних потрібно зберігати та багато разів опрацьовувати матрицю схожості, що є неприйнятним для даної предметної області; вибір міри близькості, що є непростою задачею, і є велика ймовірність виникнення помилок на даному етапі; негнучкість отриманих класифікацій. Перевага цієї групи методів у порівнянні з неієрархічними методами – їх наочність і можливість отримати детальне уявлення про структуру даних, що також є важливим для аналізу надзвичайних ситуацій на залізничному транспорті, оскільки так можливо відстежити всі процеси сполучення (розділення) кластерів, а отже можливо відслідкувати неправильні розбиття (сполучення) на (в) кластери, що є корисним для розробників, але дана особливість не є принциповою для користувачів, так як важливим є лише результат.

### *Підхід до генерації правил, за якими буде проводитись кластеризація надзвичайних ситуацій на залізниці*

Виходячи з аналізу різних методів кластеризації було з'ясовано, що оптимальним методом кластерного аналізу для розв'язання поставленої задачі є неієрархічний метод, а основні критерії які перед ним ставляться:

- висока стійкість до різних шумів та недостовірних початкових даних;
- використання мінімально можливої кількості характеристик та параметрів надзвичайної ситуації на залізничному транспорті при виконанні їх кластеризації і отримання при цьому максимально можливий якісний аналіз;
- можливість давати максимально якісний аналіз при використанні невеликої кількості попередньо кластеризованих надзвичайних ситуацій на залізничному транспорті, всі дані про які містяться у сховищі даних, але і в свою чергу можливість працювати з потужними базами даних;
- простота;
- наочність;
- висока швидкодія.

Нехай існує множина даних про надзвичайні ситуації на залізничному транспорті, які є попередньо кластеризовані. Всі ці дані збережені у сховищі даних, це означає, що вони є приведеними до єдиного формату та «очищеними». Множина даних про конкретну надзвичайну ситуацію на залізничному транспорті складається із підмножин даних про кліматичні умови, місцевість, речовину, яка перевозиться, технічні характеристики рухомого складу та ін. Після побудови кластерів, для подальшого опрацювання даних за допомогою обраного методу потрібно виявити унікальні особливості кожного із них, які будуть характерні лише для конкретної групи надзвичайних ситуацій і будуть відрізняти її від інших.

Формалізований алгоритм генерації правил кластеризації надзвичайних ситуацій на залізничному транспорті, передбачає такі основні кроки:

1. Визначити набір $X$ характеристик, який буде властивий для певного кластера надзвичайних ситуацій на залізничному транспорті.
2. Для кожної з характеристик надзвичайної ситуації ввести величину її інформативності $J_i$, числове значення якої можливо визначити за допомогою методу Information Gain Ratio[] або використовуючи досвід експертів. Але інформативність $J_i$ – це показник, який є визначений заздалегідь (тобто під час аналізу його значення практично не буде змінюватись).
3. Впорядкувати набір $X$ характеристик за ознакою зменшення їх інформативності.
4. Якщо інформативність $J_i$ певної характеристики більше середнього арифметичного інформативностей всіх характеристик $Sa_J$, включити характеристику $X_i$ до набору $X$ характеристик (рис.1).

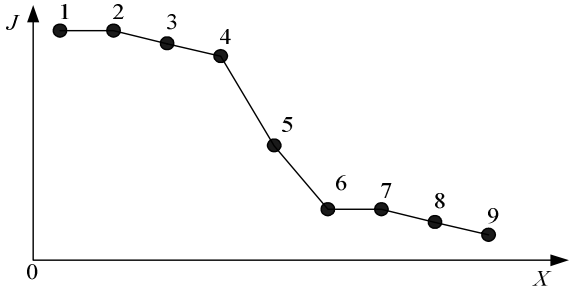


Рисунок 1 – Залежність інформативності про надзвичайну ситуацію на залізничному транспорті $J$ від її характеристики $X$

З рисунка 1 видно, що характеристики 1-5 – мають інформативність більше за середнє арифметичне всіх інформативностей, а отже характеристики 6-9 не включаються до набору $X$.

5. Для отриманого набору характеристик проводиться аналогічну процедуру для кожного з її значень $x_j$, досліджуючи інформативність

різних інтервалів значень, не включаючи у вихідну множину ті з них, значення $J_i$ у яких є меншими за середнє арифметичне інформативностей всіх значень.

6. Для отриманого набору $X$ за значеннями характеристик відібрати множину $X^I$ надзвичайних ситуації, що мають принаймні одну характеристику, числове значення якої належить визначеному інтервалу;

7. Отриману множину надзвичайних ситуацій на залізничному транспорті упорядкувати за «значущістю» $\mu$. «Значущість» кожної такої ситуації $\mu_i$ визначається як сума добутків інформативності самої характеристики та інформативності її значення, яке входить у сформований набір $X$ характеристик.

8. Якщо «значущість» $\mu_i$ надзвичайної ситуації менша середнього арифметичного $Sa_\mu$ множини $\mu$. Результатом буде шукана множина надзвичайних ситуацій для даного набору значень характеристик.

9. Аналогічно для будь-якої множини надзвичайних ситуацій будується набір ключових характеристик та їх значень.

10. Визначити правила формування кластерів з урахуванням операцій двох типів - «І» та «АБО». При цьому, показник операції «І» – величина, що визначається кількістю надзвичайних ситуацій на залізничному транспорті, значення характеристик яких лежать в визначеному діапазоні, а показник операції «АБО» визначається в такий спосіб – якщо для надзвичайної ситуації значення однієї характеристики входить у визначений діапазон, а другої – ні, але після введення певних змін або очищення даних значення другої характеристики також входить в знайдений діапазон – то це може означати, що ці характеристики мають близьке значення, наприклад атмосферний тиск та вологість повітря.

11. Побудувати дві матриці операцій, – «І» та «АБО» – та обрати з них члени з найбільшими значеннями показників операцій «І» та «АБО».

12. Сформувати новий набір ознак, які є у надзвичайних ситуацій.

13. Повторити пункти 10-12 але вже із «складними характеристиками», тобто характеристиками на вищому рівні ієрархії.

14. Сформувати правила для визначення кластерів.

Наведений алгоритм можна подати схемою, що представлена на рисунку 2.

## *Висновки*

В роботі запропоновано алгоритм розв'язання задачі кластеризації при аналізі надзвичайної ситуації на залізничному транспорті з метою її ідентифікації. Переваги алгоритму: висока швидкодія, можливість кластеризації в режимі реального часу. Недолік алгоритму: можливість появи недостовірних рішень при аналізі принципово нових надзвичайних ситуацій.

## *Література*

1. Кластерный анализ. Режим доступу: http://www.statsoft.ru/home/textbook/modules/stcluan.html#statistical.

2. Т.О. Савчук, С.І. Петришин Використання ієрархічних методів кластеризації для аналізу надзвичайних ситуацій на залізничному транспорті// Стаття, Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах» (м. Хмельницький, 2009.-№1, с.193-198).

3. Айвазян С.А., Бухштабер В.М., Енюков И.С. Прикладная статистика: Классификация и снижение размерности. – М.:Финансы и статистика, 1989. – 607 с.

4. Мандель И.Д. Кластерный анализ. – М.:Финансы и статистика, 1988. – 176с.

5. Методы анализа структуры. Режим доступу: www.sati.archaeology.nsc.rustatmethods_info.php

6.Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др. Факторный, дискриминантный и кластерный анализ: Пер. с англ. – М.: Финансы и статистика, 1989. – 215с.

7. Дюран Б., Одел П. Кластерный анализ: Пер. с англ. – М.: Статистика, 1977. – 128 с.

8. Ивахненко А.Г. Алгоритмы метода группового учета аргументов при непрерывных и бинарных признаках. Препринт. – К.: Ин-т кибернетики им. В.М.Глушкова, 1992. – 49 с.

9. Гитис П.Х. Статистическая классификация и кластерний анализ. – М.: Московский государственный горный университет, 2003. – 157 с.

10. Выбор метода кластеризации: Режим доступу:http://www.market-journal.com/ marketingovyeissledovanija/209.html

*Надійшла до редакції 30.03.2010*

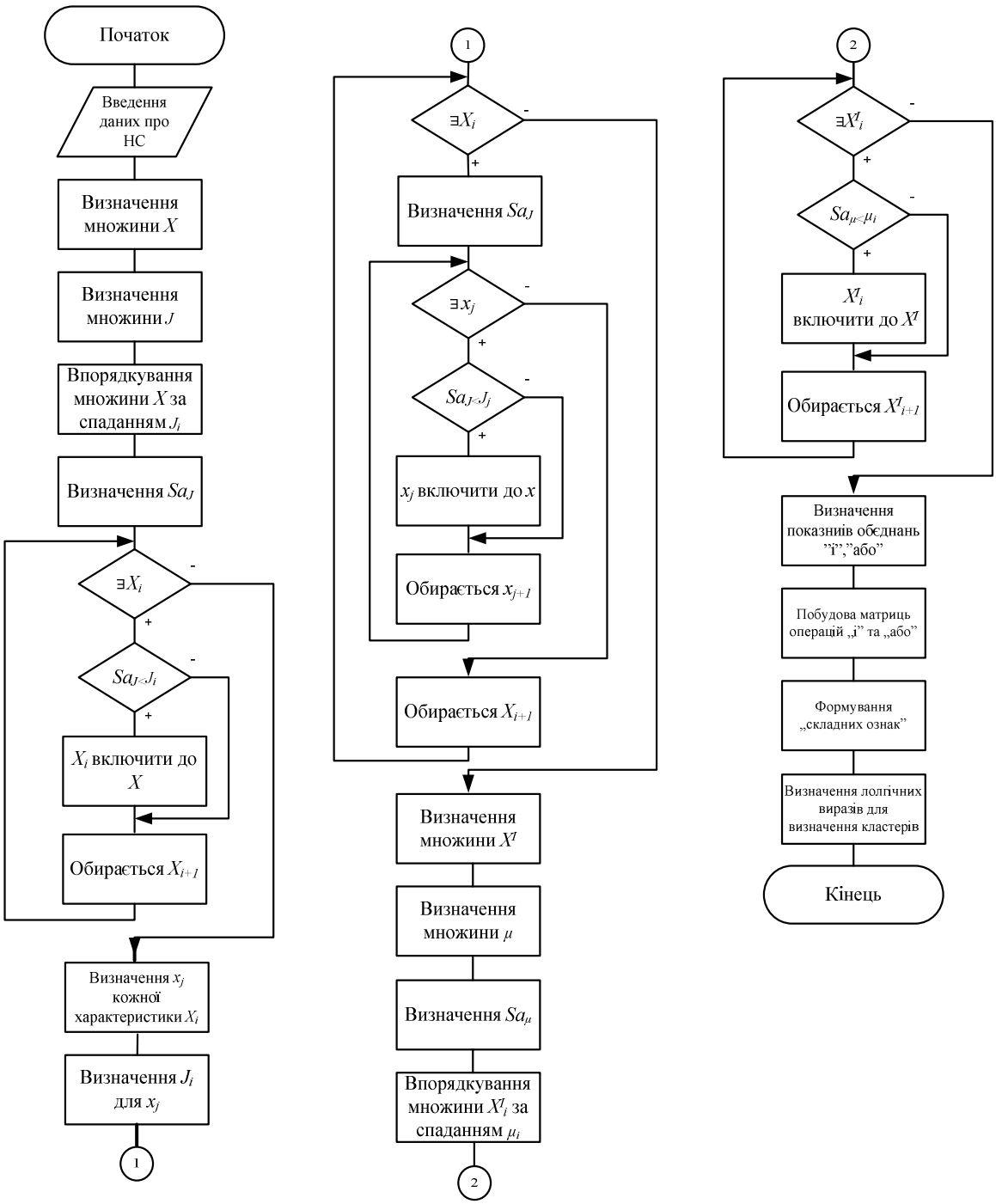


Рисунок 2 – Схема алгоритму генерації правил кластеризації надзвичайних ситуацій на залізничному транспорті