ВІДСТАНЬ ТА СТУПІНЬ БЛИЗЬКОСТІ ЯК БАЗОВІ ХАРАКТЕРИСТИКИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ НАДЗВИЧАЙНИХ СИТУАЦІЙ НА ЗАЛІЗНИЧНОМУ ТРАНСПОРТІ

Тамара Савчук[1], *Сергій Петришин*[2]

Вінницький національний технічний університет
Хмельницьке шосе, 95, Вінниця, 21021, Україна, тел.: 0432 513211, E-Mail: [2] Petrishyn@gmail.com

Анотація

В доповіді проаналізовано можливість застосування технологій Data Mining при аналізі надзвичайних ситуацій на залізничному транспорті. Формалізовано задачу кластерного аналізу, досліджено основні проблеми визначення відстані між надзвичайними ситуаціями при такому аналізі. Визначено поняття деяких мір близькості, що базуються на відстанях між надзвичайними ситуаціями на залізничному транспорті.

Вступ

На сьогоднішній день залізничний транспорт є найпоширенішим в Україні. Ним здійснюється біля 60% вантажоперевезень (включаючи перевезення шкідливих та небезпечних вантажів) по території держави [1].

Для забезпечення сталих тенденцій розвитку залізничного транспорту, що використовується для перевезення небезпечних вантажів, необхідно поєднувати технічний розвиток рухомого складу залізниць з розвиненою системою реагування на надзвичайні ситуації, які можуть виникнути під час їх перевезення. Такі системи повинні базуватись на новітніх інформаційних технологіях, що будуть використовуватись для аналізу означених надзвичайних ситуацій, з метою запобігання/зменшення їх виникнення/можливих наслідків.

Аналіз існуючих методів аналізу надзвичайних ситуацій на залізничному транспорті

Велику кількість задач, зокрема і задачу аналізу надзвичайних ситуацій на залізничному транспорті, допомагають розв'язати технології Data Mining в залежності від задач, які розв'язуються (описові задачі та задачі прогнозування). Всі алгоритми аналізу даних поділяють на supervised learning (навчання з учителем) та unsupervised learning (навчання без учителя). В першому випадку задача аналізу розв'язується в декілька етапів. Спочатку за допомогою певного алгоритму будується модель даних, що аналізуються. Потім ця модель навчається на навчальних вибірках до того моменту, поки вона не почне працювати коректно. Unsupervised learning використовується коли немає ніяких попередніх знань про дані, що аналізуються. Основними із задач Data Mining є такі: класифікація, регресія, пошук асоціативних правил і кластеризація [2].

Постановка задачі

Перед розв'язанням задачі кластеризації для аналізу надзвичайних ситуацій на залізничному транспорті потрібно 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'язки 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.

Нехай Y – матриця, в якій кожен стовпчик $\{y_{i1},...,y_{ij},...y_{im}\}$ описує певну надзвичайну ситуацію, тобто y_{ij} – певна характеристика окремої надзвичайної ситуації.

$$Y = \{Y_1, Y_2, ..., Y_n\} = \begin{Bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{Bmatrix},$$

де Y_i – конкретна надзвичайна ситуація на залізничному транспорті;

y_{ij} – значення конкретного j-го параметру i-ї надзвичайної ситуації;

m – кількість параметрів надзвичайних ситуацій, що збережені в базі даних.

Отже, на основі зазначеного можливо сформулювати таку задачу:

Є множина надзвичайних ситуацій на залізничному транспорті $Y = \{Y_i\}\,(i = \overline{1,n})$ (дані про еталонні надзвичайні ситуації зберігаються в сховищі даних, тобто всі значення y_{ij} 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 несхожими [3].

Відстані та ступені близькості надзвичайних ситуацій на залізничному транспорті

Складності у формалізації задачі кластерного аналізу надзвичайних ситуацій на залізничному транспорті пов'язані з визначенням поняття їх однорідності [4] та слабкою структурованістю даних.

В загальному випадку однорідність двох i-ї та j-ї надзвичайної ситуацій на залізничному транспорті визначається завданням правила обрахунку величини ψ_{ij}, що характеризує або відстань $a(Y_i, Y_j)$ між об'єктами Y_i та Y_j із досліджуваної множини надзвичайних ситуацій на залізничному транспорті $Y = \{Y_i\}\,(i = \overline{1,n})$ або ступінь близькості $\omega(Y_i, Y_j)$ між тими ж ситуаціями. Якщо задана функція $a(Y_i, Y_j)$, то близькі за значенням цієї метрики надзвичайні ситуації вважаються однорідними, тобто такими, що належать одному кластеру. Але при цьому необхідно порівнювати $a(Y_i, Y_j)$ з певними пороговими значеннями, що визначаються в кожному випадку. Означений підхід доцільно використовувати для визначенні міри близькості $\omega(Y_i, Y_j)$ при формуванні однорідних кластерів надзвичайних ситуацій на залізничному транспорті. При цьому повинні бути дотримані такі вимоги:

- вимога симетрії ($\omega(Y_i, Y_j) = \omega(Y_j, Y_i)$);

- вимога максимальної подібності надзвичайних ситуацій самих з собою ($\omega(Y_i, Y_i) = \max(\omega(Y_i, Y_j))$);

- вимога відповідності між відстанню між надзвичайними ситуаціями на залізничному транспорті та мірою близькості між ними(якщо $a(Y_1, Y_2) \geq a(Y_2, Y_3)$ то $\omega(Y_1, Y_2) \leq \omega(Y_2, Y_3)$).

Вимірювання відстані між надзвичайними ситуаціями

Відстанню між надзвичайними ситуаціями Y_i та Y_j або метрикою називається невід'ємна дійсна функція $a(Y_i, Y_j)$, якщо [5]:

- $a(Y_i, Y_j) \geq 0$ для всіх Y_i та Y_j з множини $Y = \{Y_i\}\,(i = \overline{1,n})$;

- $a(Y_i, Y_j) = 0$ тоді і тільки тоді, коли $Y_i = Y_j$;

- $a(Y_i, Y_j) = a(Y_j, Y_i)$;

- $a(Y_i, Y_j) \leq a(Y_i, Y_k) + a(Y_k, Y_j)$, де Y_i, Y_j та Y_k – будь-які три надзвичайні ситуації на залізничному транспорті з множини $Y = \{Y_i\}\,(i = \overline{1,n})$.

При кластерному аналізі надзвичайних ситуацій виникає проблема вимірювання відстані між окремими такими ситуаціями. Основні труднощі, що виникають при цьому [5]:

- неоднозначність вибору способу нормування;
- неоднозначність визначення відстані між об'єктами.

Міри близькості, що базуються на відстанях

Відстані між надзвичайними ситуаціями на залізничному транспорті передбачають їх представлення у вигляді точок m-вимірного простору.

Евклідова відстань [6] є одним з найбільш використовуваних метрик в кластерному аналізі, оскільки вона відповідає інтуїтивним уявленням про близькість і своєю квадратичною формою відповідає класичним статистичним конструкціям. Геометрично дану метрику доцільно використовувати для об'єднання об'єктів в кулястих скупченнях, які є типовими для слабко корельованих множин.

Формула загальної евклідової відстані має вигляд:

$$a_E(Y_i,Y_j)=\sqrt{(y_{i1}-y_{j1})^2+(y_{i2}-y_{j2})^2+...+(y_{im}-y_{jm})^2},$$

де $a_E(Y_i,Y_j)$ – евклідова відстань між двома надзвичайними ситуаціями на залізничному транспорті Y_i та Y_j;

$y_{i1}, y_{i2}, \ldots, y_{im}$ – вектор значень характеристик, що описує i-ту надзвичайну ситуацію на залізничному транспорті;

$y_{j1}, y_{j2}, \ldots, y_{jm}$ – вектор значень характеристик, що описує j-ту надзвичайну ситуацію на залізничному транспорті.

Дану метрику доцільно застосовувати в таких випадках:

- значення параметрів $y_{i1}, y_{i2}, \ldots, y_{im}$ однорідні за фізичним змістом, і якщо встановлено, що всі вони однаково важливі з точки зору розв'язку задачі про віднесення надзвичайної ситуації на залізничному транспорті до певного кластера;
- простір ознак співпадає з геометричним простором дійсності і поняття близькості надзвичайних ситуацій співпадає з поняттям геометричної близькості в цьому просторі.

Відстань за Хемінгом є середнім різниць по координатах. В більшості випадків ця міра близькості приводить до аналогічних результатів як і евклідова відстань, але для неї вплив великих викидів зменшується оскільки вони не підносяться до квадрату.

Загальний вигляд формули відстані за Хемінгом має вигляд:

$$a_H(Y_i,Y_j)=|y_{i1}-y_{j1}|+|y_{i2}-y_{j2}|+...+|y_{im}-y_{jm}|,$$

де $a_H(Y_i,Y_j)$ – відстань за Хемінгом між двома надзвичайними ситуаціями на залізничному транспорті Y_i та Y_j;

$y_{i1}, y_{i2}, \ldots, y_{im}$ – вектор значень характеристик, що описує i-ту надзвичайну ситуацію на залізничному транспорті;

$y_{j1}, y_{j2}, \ldots, y_{jm}$ – вектор значень характеристик, що описує j-ту надзвичайну ситуацію на залізничному транспорті.

Пікова відстань припускає незалежність між випадковими змінними, що говорить про відстань в ортогональному просторі. Але в практичних додатках ці змінні не є незалежними.

Формули пікової відстані має вигляд:

$$a_L(Y_i,Y_j)=\frac{1}{m}\cdot\left(\frac{|y_{i1}-y_{j1}|}{y_{i1}+y_{j1}}+\frac{|y_{i2}-y_{j2}|}{y_{i2}+y_{j2}}+...+\frac{|y_{im}-y_{jm}|}{y_{im}+y_{jm}}\right),$$

де $a_L(Y_i, Y_j)$ – пікова відстань між двома надзвичайними ситуаціями на залізничному транспорті Y_i та Y_j;

$y_{i1}, y_{i2}, \ldots, y_{im}$ – вектор значень характеристик, що описує i-ту надзвичайну ситуацію на залізничному транспорті;

$y_{j1}, y_{j2}, \ldots, y_{jm}$ – вектор значень характеристик, що описує j-ту надзвичайну ситуацію на залізничному транспорті.

Висновки

Таким чином, в роботі визначено актуальність проблеми аналізу надзвичайних ситуацій на залізничному транспорті, проаналізовано можливість застосування технологій Data Mining при такому аналізі. Для аналізу надзвичайних ситуацій на залізничному транспорті доцільно використовувати кластеризацію, як таку, що характеризується ітераційним пошуком оптимального рішення; можливістю вибору інформативних ознак та мір схожості двома об'єктами, об'єктом і кластером, двома кластерами; побудовою науково обгрунтованої класифікації багатовимірних спостережень на підставі сукупності відібраних показників та виявлення внутрішніх зв'язків між надзвичайними ситуаціями на залізниці, що аналізуються.

Література:

[1] 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).

[2] 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.

[3] Т.О. Савчук, С.І. Петришин Порівняльний аналіз використання методів кластеризації для ідентифікації надзвичайних ситуацій на залізничному транспорті// Стаття, Наукові праці Донецького національного технічного університету. – Серія «Інформатика, кібернетика і обчислювальна техніка». – 2010. – Випуск 11(134). – С. 135-141.

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

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

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