Національного технічного університету України Київський політехнічний інститут. 2013. № 2. С. 39-47.

7. Жора В. Підхід до моделювання рольової політики безпеки. *Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні*. 2003. Вип. 7. С.45-49.

8. Девянин П. Н. Модели безопасности компьютерных систем : учеб. пособие для студ. высш. учеб. заведений. М. : Издательский центр «Академия», 2005. 144 с.

9. Бабенко Л. К., Басан А. С., Макаревич О. Б. Мандатный доступ в СУБД. *Известия Южного федерального университета*. Технические науки. 2005. № 4 (48). С. 133-136.

10. Баранов П. А. Проблемы реализации мандатного доступа (модель Белла-ЛаПадулы) к ресурсам вычислительных систем. *Проблемы информационной безопасности. Компьютерные системы* 2005. № 1. С. 7-14.

11. Яковів І., Корнейко О., Черноног О. Класифікація і аналіз моделей систем захисту інформації в комп'ютерних системах. *Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні* : науково-технічний збірник. 2001. Вип. 3. С. 71-76.

12. Кононович І. В. Динаміка кількості інцидентів інформаційної безпеки. *Інформатика та математичні методи в моделюванні*. 2014. № 1. С. 35-43.

Одержано 17.10.2018

УДК 343.346.8:004.056.53

Володимир Михайлович СТРУКОВ,
кандидат технічних наук, доцент
завідуючий кафедри інформаційних технологій
факультету № 4 (кіберполіції)
Харківського національного університету внутрішніх справ;
ORCID: https://orcid.org/0000-0003-4722-3159;
Дмитро Юрійович УЗЛОВ,
кандидат технічних наук,
начальник Управління інформаційно-аналітичної підтримки
ГУНП в Харківській області;
ORCID: https://orcid.org/0000-0003-3308-424X;
Олексій В'ячеславович ВЛАСОВ,
заступник начальника Управління інформаційно-аналітичної підтримки
ГУНП в Харківській області

МАКСИМІННИЙ КРИТЕРІЙ ВИЗНАЧЕННЯ ПЕРВИННИХ ЦЕНТРІВ В АЛГОРИТМАХ КЛАСТЕРИЗАЦІЇ K-MEANS

В базах даних органів внутрішніх справ на поточний момент накопичено десятки мільйонів записів кримінально значимої інформації. І кожного дня кількість цих даних зростає. Результати

запитів до системи іноді містять кілька сотень або навіть тисяч записів, які в подальшому необхідно обробляти вручну. Крім того, необхідно враховувати такі особливості масивів даних ОВС:

1) велика кількість ознак, що характеризують об'єкти (до сотні ознак);

2) різна природа ознак (як правило, нечислових), вимірюваних в різних шкалах;

3) можливість наявності декількох записів про одне й те ж об'єкті (відома проблема «двійників»);

4) можливість наявності пропусків (відсутність значень там, де вони повинні знаходитися) у масивах даних в силу ряду суб'єктивних і об'єктивних причин.

Очевидно, що аналізувати результати таких запитів вручну, як це робиться в більшості випадків сьогодні, вкрай неефективно.

Ефективним механізмом обробки надвеликих даних є технології Data Mining, Text Mining, Web Mining. В основі цих наукоємних технологій лежать методи класифікації-кластеризації. Одними з простих та широко використовуваних алгоритмів кластеризації є алгоритми K-means і C-means [1]. Їх основними перевагами є простота реалізації, результативність і висока швидкодія. Головний недолік – заздалегідь задана кількість кластерів, на які необхідно розбити вихідний набір об'єктів. Крім того, в алгоритмах цього класу на першій стадії при визначенні первинних центрів шуканих кластерів, як правило, випадковим чином обираються K об'єктів. Однак у випадках, коли кількість досліджуваних об'єктів досягає сотень тисяч і мільйонів записів, такий підхід є завідомо неприйнятним. У цьому випадку є доцільним мати досить «хорошу» початкову вибірку первинних центрів, яка в подальшому забезпечить рішення, близьке до локального (або глобального) екстремуму. В роботах [2, 3] запропонований критерій вибору первинних центрів і алгоритм кластеризації, заснований на його використанні. Його ефективність значно вище алгоритмів, заснованих на випадковому виборі первинних центрів. Однак подальше експериментальне дослідження цього алгоритму виявило випадки, коли він дає неточні результати, хоча і близькі до локальних екстремумів.

В даній роботі пропонується максимінний критерій вибору первинних центрів кластерів в алгоритмах K-means. Його суть полягає у наступному.

Нехай $X=\{x_{ij}\} \in R^{mxn}$ – матриця «об’єкт-властивість», де x_{ij} – значення j-ї властивості i-го об’єкта a, i=1,2, …, m; j=1,2, …, n, при цьому кожний об’єкт може бути описаний вектором-рядком

властивостей $x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \in R^n$. Вважатимемо, що значення всіх n ознак задані у звичайній числовій шкалі, хоча, як показано в [4], до такого випадку можна звести і ситуацію, коли значення ознак об'єктів задані в інших шкалах.

K – кількість кластерів у задачі;

$X^c=\{x_i^c\}$, $i=1, 2, \ldots, K$ – множина шуканих центрів кластерів.

Нехай також у просторі R^n об'єктів задана відстань між ними у звичайній евклідовій метриці:

$$d_{il} = \sqrt{\sum_{j=1}^{n} (x_{ij} - x_{lj})^2}. \quad (1)$$

В якості цільової функції у задачі пошуку первинних центрів К кластерів запропонована наступна:

$$d^c = \underset{i}{Max} \underset{j}{Min} \; d_{ij} \quad (2)$$

де d_{ij} – це відстань від i-тої точки до j-того центра, $i = 1, 2, \ldots, m$; j=1,2, …, K,.

Експериментальне дослідження програмної реалізації запропонованого критерію на тестових прикладах показало, що він ефективніше критерію, запропонованого в роботі [3], в середньому на 10%.

Список бібліографічних посилань

1. Aggarwal C. C., Reddy C. K. Data clustering. *Algorithms and Applications*. Cham : Springer Ltd. Publ., Switzerland, 2015. 734 p.

2. Струков В. М. Модифікація алгоритму кластеризації K-means // Матеріали Всеукраїнської наук.-практ. конф. «Актуальні питання протидії кіберзлочинності та торгівлі людьми» (15.11.2017 р., м. Харків). Х. : ХНУВС, 2017. С. 162-164.

3. Струков В. М. Критерій і алгоритм вибору первинних центрів кластерів в алгоритмі кластеризації K-means // Тези доповідей IX Міжнародної науково-технічної конференції «Інформаційно-комп'ютерні технології – 2018» (20-21 квітня 2018 р.). Житомир : Вид. О. О. Євенок, 2018. С. 49. URL: https://conf.ztu.edu.ua/wp-content/uploads/2018/05/49.pdf (дата звернення: 31.10.2018).

4. Бодянский Е. В., Струков В. М., Узлов Д. Ю. Задача оценки близости многомерных объектов анализа данных. *Управляющие системы и машины*. 2016. № 6. С. 67-72.

Одержано 03.11.2018