УДК 004.45:004.9

В. В. Юзефович
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Вирішення задачі групування рухомих об'єктів у системах моніторингу з використанням методів кластеризації

Розглянуто підхід до вирішення задачі автоматичного групування рухомих об'єктів на основі використання модифікованого методу «найближчого сусіда», який відрізняється від відомого аналізом відстаней від поточного елемента множини до границь існуючих кластерів.

Ключові слова: *система моніторингу, рухомий об'єкт, групування, кластеризація, кластер.*

Вступ

Для вирішення задач спостереження за рухомими об'єктами (РО) та контролю їхніх дій у певній частині простору створюються спеціальні системи моніторингу, що поєднують у собі засоби добування, збору, обробки та розподілу інформації про РО між споживачами.

Сучасні умови функціонування таких систем характеризуються зростаючими обсягами наявних у районі їхньої відповідальності об'єктів спостереження, збільшенням кількості різноманітних задач, що покладаються на систему моніторингу з одночасною вимогою щодо зниження часу «знаходження» даних про об'єкти в контурах обробки. У результаті, в зазначених інформаційних системах можуть виникати перенавантаження та, як наслідок, втрати інформації.

Разом з тим, згідно із принципом агрегування інформації у ієрархічних системах [1], для прийняття рішень на більшості рівнів управління достатньо своєчасно отримувати певним чином узагальнену (без зайвих подробиць) інформацію.

Отже, одним із шляхів зменшення інформаційного навантаження, як на системи моніторингу та засоби передавання інформації, так і на споживачів, може бути групування одиночних РО, що спостерігаються системою, яке полягає в об'єднанні об'єктів зі схожими параметрами у групи (кількість яких може бути значно менша за кількість окремих РО) та подальше оперування груповими об'єктами [2].

Таке «стискання» інформації забезпечить зменшення об'ємів даних, що передаються споживачам і, відповідно, збільшить швидкість передачі необхідних масивів інформації.

Постановка задачі

Для вирішення задачі групування об'єктів (даних про об'єкти) використовуються різні методи кластеризації. При цьому загальною характерною рисою всіх реалізованих і більшості відомих способів групування повітряних, наземних і надводних РО в просторі є використання просторових стробів правильної геометричної форми, розміри яких або задаються апріорі (на основі аналізу тактики дій противника під час навчань або локальних конфліктів), або ж групи визначаються автоматично відповідно до параметрів, що задані оператором [2]. Такими параметрами є кількість групових об'єктів у зоні відповідальності, або порогове значення відстані між групами (найменша допустима відстань між найближчими об'єктами різних груп, або максимальна відстань між сусідніми об'єктами всередині однієї групи).

Використання стробів із визначеними розмірами виключає врахування будь-яких особливостей поточної обстановки. Недоліком інших підходів є складність (особливо в умовах великої кількості РО) оперативного визначення значень або порогової відстані між групами (об'єктами в групі), або кількості груп. Оскільки ж положення РО з кожним циклом оновлення обстановки змінюються, може виникати задача постійного уточнення вказаних параметрів. Фактично використання зазначених підходів потребує постійної активної участі людини-оператора, а завдання, які на неї покладаються, не є елементарними і потребують суттєвих навичок і знань.

Отже, виникає актуальна задача здійснення автоматичного групування об'єктів (без участі оператора у самому процесі групування), що є метою даної роботи.

Зміст дослідження

Ієрархічна агломеративна кластеризація є одним із найбільш поширених методів евристичної кластеризації, суть якої полягає у побудові дерева зв'язків (дендрограми), починаючи з припущення, що всі елементи множини є окремим кластером. Для цього попередньо, як і при використанні будь-якого іншого методу, розраховується симетрична квадратна матриця відстаней D між об'єктами спостереження, що утворюють множину РО $X = \{x_1, \ldots, x_p\}$:

$$D = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{p1} & d_{p2} & \cdots & 0 \end{pmatrix},$$

де d_{ij} — відстань між i-м та j-м елементами множини X, що визначається за одним із відомих виразів [3], ($d_{ij} = d_{ji}$).

Після побудови дерева зв'язків у класичному випадку оператор має визначити (задати) порогову відстань (кількість кластерів) для «розрізання» дендрограми й отримання груп РО.

Існують різні методи побудови дерева зв'язків, що жодним чином не залежать від способу отримання самих відстаней d між елементами множини X. Однак

результати побудови дендрограми суттєво залежать від обраного методу. На рис. 1 показані різні можливі розподіли деяких об'єктів у просторі їхніх параметрів. Аналіз рисунку показує ряд проблем вирішення задачі їхнього розбиття за бажаними (з точки зору дослідника, або відповідно до фізики процесу) кластерами. Наприклад, відстань між окремими об'єктами кластера *C* більше ніж відстань між кластерами (об'єктами кластерів) *B* та *C*; середні значення параметрів об'єктів у кластерах *E* і *F* та *K* і *H* однакові (центри кластерів співпадають); кластери *P* і *Q* з'єднані ланцюгом об'єктів. Зазначені та інші ситуації і чинники призвели до виникнення великої кількості різноманітних методів кластеризації, жоден з яких не є універсальним. У роботі [3] також наведено досить широкий аналіз різних груп методів кластеризації із зазначенням їхніх недоліків і переваг, які, в свою чергу, визначаються конкретною прикладною задачею. Вагомою, з точки зору вибору метода кластеризації, особливістю задачі групування саме РО є типовість ситуацій групування, що на рис. 1 показані кластерами *A*, *B* і *C* та *E* і *F*. Наприклад, колона техніки (кластер *C*), або, для військових систем, РО безпосереднього прикриття іншого об'єкта та РО лінії оборони (кластери *E* і *F*). Більшість відомих методів кластеризації «не впораються» із задачею виділення таких кластерів, оскільки «намагаються» скупчити елементи навколо деяких центрів кластерів у різних математичних сенсах [3].

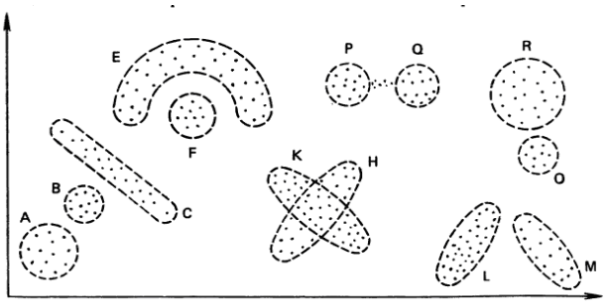

Рис. 1. Приклади розподілу елементів множин і кластерів, отриманих різними методами (зображення запозичене з [3])

З метою перевірки останнього твердження щодо методів кластеризації здійснено модельний експеримент за допомогою програмного стенду, розробленого для дослідження можливості застосування різних методів кластеризації для вирішення задачі групування РО, а також вивчення особливостей різних методів. На рис. 2 показано вхідний розподіл елементів, що підлягають групуванню та результат їхнього розподілу за кластерами, отриманий за однією з можливих агломеративних процедур кластеризації, яка поєднує елементи, найближчі до деяких центрів кластерів, положення яких визначається та уточнюється в процесі групування. Вхідний розподіл елементів (рис. 2,*а*) дозволяє достатньо просто візуально виділити два ланцюги об'єктів (за типом кластера *C* на рис. 1) та, можливо, один окремий об'єкт. Од-

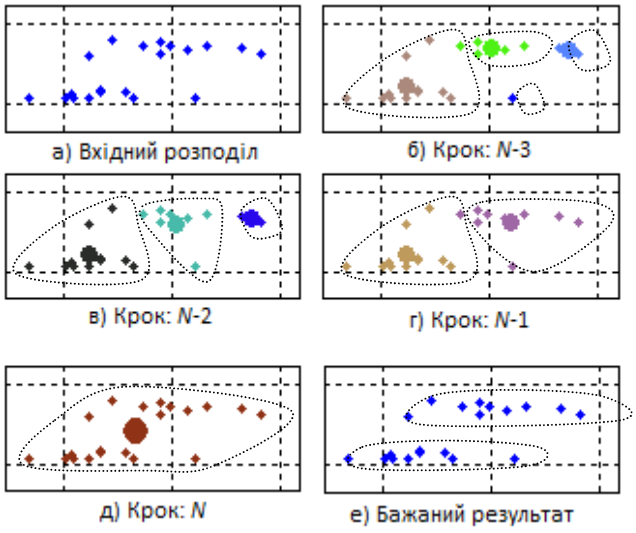

Рис. 2. Результат кластеризації із приєднанням елементів, найближчих до центра кластера

нак з рис. 2,*б* – 2,*д*, де показані декілька останніх кроків зазначеної процедури кластеризації, видно, що виділення ланцюгів у такому випадку дійсно є неможливим. (На рис. 2,*е* показано бажаний результат кластеризації).

Аналіз зображених на рис. 2 розподілів та інших можливих розподілів РО дозволяє зазначити, що для вирішення задачі їхнього групування необхідно використовувати підходи, що дозволяють приєднувати до кластера елементи, які знаходяться найближче до його границь незалежно від форми кластера. Крім того, для автоматичного вирішення задачі отримання розподілу об'єктів (без участі оператора) в алгоритм групування РО необхідно закласти деяку логіку (математичну процедуру) зупинки процесу кластеризації. Причому, для практики важливо, щоб сама процедура була максимально простою та видавала результати, які б легко піддавалися трактуванню, а отже, за необхідності, забезпечувала би просту процедуру впливу на результати кластеризації у разі неотримання бажаного розподілу об'єктів.

Виходячи з аналізу методів кластеризації, наведеного у [3], за основу для вирішення задачі групування РО обрано метод «найближчого сусіда», або «одиночного зв'язку». При цьому, для забезпечення бажаного результату при групуванні елементів, розподілених за типом «ланцюг» (кластер *C* на рис. 1), будемо розраховувати, на відміну від класичного підходу, не відстані між центрами існуючих кластерів, а чергову мінімальну відстань між елементами множини *X*, що можуть на момент розгляду відноситися до певного кластера, або бути «вільними».

Отже, в нашому випадку, процедура кластеризації буде наступною.

1. На початку процедури кластеризації кожен елемент множини *X*, із загальною кількістю елементів *N* вважається окремим кластером. Першим кроком є визначення номерів пари елементів (*i*, *j*), де $i \neq j$, матриці відстаней *D*, розміром $N \times N$, для яких відстань d_{ij} є мінімальною. Ці елементи об'єднуються в один кластер, а відстань d_{ij} (та d_{ji} відповідно), з подальшого аналізу виключається. Загальна кількість кластерів зменшується на одиницю.

2. Розраховується центр кластера, наприклад, для просторового групування як середнє арифметичне значень параметрів елементів *i* та *j*, за якими визначалася відстань між ними при формуванні матриці *D*. Формується список елементів кластера.

3. Визначається чергова пара елементів матриці *D* з мінімальною відстанню з подальшим застосуванням одного із наступних правил:

а) якщо обидва елементи на час їхнього аналізу не відносяться до жодного з існуючих кластерів — вони утворюють новий кластер, для якого здійснюються операції пункту 2. Кількість кластерів зменшується на одиницю.

б) якщо один із елементів уже знаходиться у списку одного із кластерів — другий елемент додається до цього списку. Кількість кластерів не змінюється. Для даного кластера розраховується новий центр та уточнюється список його елементів.

в) якщо обидва елементи на час їхнього аналізу відносяться до різних кластерів, ці кластери поєднуються, а всі їхні елементи заносяться до одного списку. Кількість кластерів зменшується на одиницю та розраховується новий центр утвореного кластера.

г) якщо обидва елементи відносяться до одного й того ж самого кластера — даний крок кластеризації вважається «холостим». Відстань d_{ij} (d_{ji}) між цими еле-

ментами виключається з подальшого аналізу. Кількість кластерів і координати їхніх центрів не змінюються.

4. Агломеративна процедура закінчується, коли на черговому кроці всі елементи множини поєднаються в один кластер (тобто буде отримане повне дерево зв'язків).

Як видно із опису процедури, центри кластерів не «приймають участі» у розрахунках і необхідні лише для візуального відображення кластерів (груп). Тому спосіб розрахунку центрів кластерів з точки зору процедури кластеризації є некритичним.

Для перевірки дієздатності запропонованого підходу та можливості групування скупчень елементів типу «ланцюг» було розроблено його математичну модель у середовищі MatLab. На рис. 3 показано той самий розподіл елементів, що і на рис. 2, та результати групування елементів за запропонованою процедурою. Як видно з рис. 3, даний підхід, як і передбачалося, дозволяє «виявляти» скупчення об'єктів типу «ланцюг».

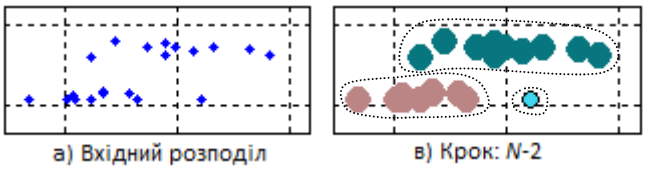

Рис. 3. Результат кластеризації скупчень елементів типу «ланцюг» за запропонованою процедурою

Наступним важливим кроком процедури є визначення, на якому кроці кластеризації необхідно «зупинитися» для отримання «бажаних» результатів.

Простими «штучними» підходами до вирішення цієї задачі, як вже зазначалося, є апріорне визначення бажаної кількості кластерів — k, або задавання порогової відстані ($d_{\max}$) між кластерами. В обох випадках відсутня необхідність побудови повного дерева зв'язків, однак їм притаманні недоліки, зазначені вище.

При викладені запропонованого підходу до групування РО зазначалася можливість виникнення «холостих» кроків кластеризації у випадку, коли обидва елементи на момент їхнього розгляду входять до одного кластера (4 крок процедури, пункт «г»), і на цьому кроці ніяких змін ні у кількості кластерів, ні у параметрах окремого кластера не відбувається. На перший погляд така ситуація є недоліком запропонованої процедури, оскільки призводить до обчислювальних витрат, без будь-яких наслідків. Однак можна припустити, що повторення такої ситуації протягом декількох кроків підряд свідчить про певну «стабільність», сталість поточних результатів розбиття множини. При цьому, вочевидь, чим більшою кількістю «холостих» кроків характеризується перехід від i кластерів до $i - 1$, тим стабільнішим є результат. Отже, зберігання кількості «холостих» кроків процедури — h між кожною зміною кількості кластерів дозволить визначити найбільш стабільні розподіли об'єктів між кластерами. Однак практичні дослідження показали, що чим вище (за ієрархією) по дереву зв'язків ми «рухаємося», тим потенційно більша кількість «холостих» кроків (N_{XK}) буде спостерігатися. Наприклад, у наведеній нижче таблиці показано кількість таких кроків для різної кількості утворених кластерів (N_K) (у даному модельному експерименті розглядалася множина X з 50-ти елементів, скупчених навколо 5-ти центрів розсіяння) в процесі кластеризації.

Розподіл «холостих» кроків кластеризації на множині кластерів

$N_К$	$N_{ХК}$	$N_К$	$N_{ХК}$	$N_К$	$N_{ХК}$	$N_К$	$N_{ХК}$	$N_К$	$N_{ХК}$	$N_К$	$N_{ХК}$	$N_К$	$N_{ХК}$
50–38	**0**	30–29	**0**	24	**0**	19–16	**3**	12	**5**	**8**	**14**	4	**20**
37	**1**	28	**1**	23	**2**	15	**4**	11	**9**	7	**4**	3	**135**
36–32	**0**	27–26	**0**	22–21	**1**	14	**1**	10	**2**	6	**21**	2	**34**
1	**1**	25	**2**	20	**0**	13	**7**	9	**0**	5	**38**	1	**84**

Як результат, максимальне значення кількості «холостих» кроків прагне верхніх ієрархічних рівнів дендрограми.

Більш ефективним для визначення бажаного числа кластерів може бути використання виразу для розрахунку середньозваженого значення величини де як «вага» кожного значення кількості кластерів використовується його «частотність» — кількість «холостих» ходів h.

Таким чином, кількість кластерів k_o для заданої множини РО буде розраховуватися за виразом:

$$k_o = round\left(\sum_{i=1}^{X} h_i \cdot i \Big/ \sum_{i=1}^{X} h_i\right),$$

де $round(\cdot)$ — округлення до найближчого цілого; i — кількість кластерів при якій спостерігалось h_i «холостих» кроків.

Запропонований підхід до визначення кількості кластерів було перевірено шляхом моделювання. На рис. 4 показано приклад моделювання автоматичного розподілу об’єктів по кластерам. На рис. 4,*а* — результат автоматичної кластеризації із кількістю кластерів, що дорівнює середньозваженому значенню, отриманому за останнім виразом. На рис. 4,*б*) – 4,*г*) показані інші найбільш «стійкі» (за напрямком декомпозиції) ситуації, що відповідають іншим більшим значенням h_i.

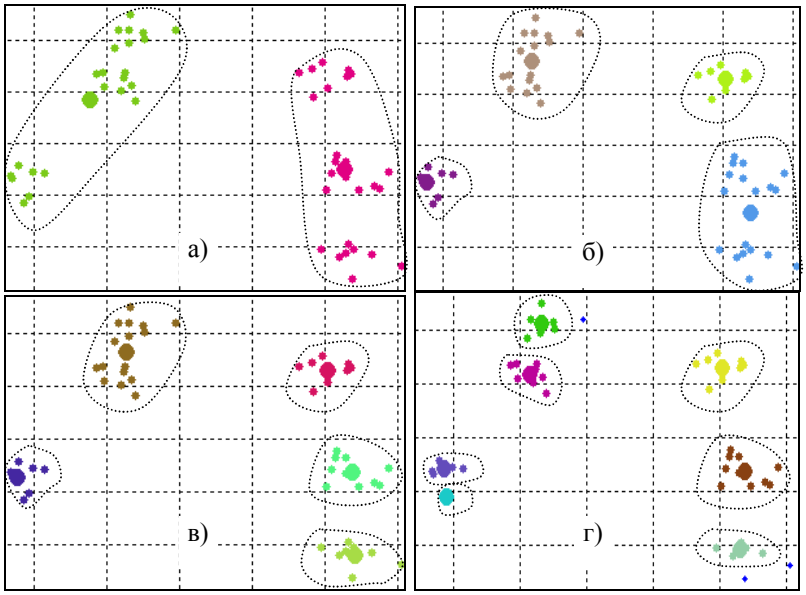

Рис. 4. Результати групування скупчень елементів
за запропонованою процедурою

Як видно з рис. 4, при використанні запропонованої процедури кластеризації (групування) доцільно надати оператору можливість застосовувати команди «деталізувати» (здійснити декомпозицію), або «узагальнити» обстановку.

Висновки

Запропоновано нову процедуру групування РО на основі методу кластеризації — «найближчого сусіда». В даній процедурі, на відміну від класичного вирішення, при формуванні кластерів не використовуються їхні центри, а розраховується відстань від поточного елемента множини до границь існуючих кластерів. Крім того, запропоновано модифікацію методу «найближчого сусіда», яка полягає у збереженні та аналізі «холостих» кроків кластеризації для визначення стійких станів (стійких результатів кластеризації). Найбільш стійку кількість кластерів (груп об'єктів) запропоновано визначати за середньозваженим значенням кількості «холостих» кроків.

Перевагами запропонованого підходу є його аналітична простота та простота практичної реалізації, можливість отримання результатів кластеризації автоматично (без участі оператора), незалежність результатів кластеризації від початку та порядку аналізу об'єктів кластеризації. Очевидно, що даний підхід не відкидає застосування для зупинки процесу кластеризації апріорно заданої кількості потрібних кластерів або максимально допустимої відстані.

До недоліків даного підходу можна віднести геометричне зростання кількості «холостих» кроків процедури кластеризації при збільшенні числа об'єктів, які підлягають групуванню, що у відповідній прогресії збільшує обчислювальні витрати та час отримання результатів.

1. *Новиков Д.А.* Механизмы функционирования многоуровневых организационных систем / Д.А. Новиков. — М.: Фонд «Проблемы управления», 1999. — 161 с.

2. *Кореньков В.* Агрегирование информации — эффективный способ борьбы с информационными перегрузками / В. Кореньков, П. Моисеенко, С. Семенов // Воздушно-космическая оборона. — 2006. — № 3 (28).

3. *Мандель И.Д.* Кластерный анализ / И.Д. Мандель. — М.: Финансы и статистика. — 1988. — 176 с.

Надійшла до редакції 20.11.2015