ІНФОРМАЦІЙНІ РЕСУРСИ ТА СИСТЕМИ
INFORMATION RESOURCES AND SYSTEMS

Anatolii V. Kuzmin[1], PhD (Physics and Mathematics), Associate professor, Faculty of Computer Science and Cybernetics
ORCID 0000-0001-5439-6387 ***e-mail***: kuzmin_a_b@ukr.net

Leonid D. Grekov[2], Doctor of technical sciences, Director of SSPC "Pryroda"
ORCID 0000-0002-1604-7730

Nataliia M. Kuzmina[3], PhD (Physics and Mathematics), Associate professor, Faculty of Informatics
ORCID 0000-0003-0136-1441 ***e-mail***: n.m.kuzmina@npu.edu.ua

Oleksii A. Petrov[4], PhD in Geography and GIS
ORCID 0000-0001-9828-2007

Olena M. Medvedenko[4], Director SPE "Agroresurssystems" LLC
ORCID 0000-0002-9178-9638

[1]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
[2]SSPC "Pryroda", Kyiv, Ukraine
[3]National Pedagogical Dragomanov University, Kyiv, Ukraine
[4]SPE "Agroresurssystems", Kyiv, Ukraine

COMPUTATIONAL PROCEDURES FOR THEMATIC PROCESSING OF SPACE IMAGERY FOR AGRICULTURAL RESOURCES MONITORING (PART 2)

Abstract. *The universal fast algorithm of cluster analysis is considered. The proposed algorithm is a grid type, it uses the point density parameter in the grid cell and the ratio between neighborhoods to unite of neighboring dense cells into clusters.*

The algorithm sequentially calculates for each point the number of the cell to which it belongs, then generates groups of points for each non-empty cell. Then it sequentially unites cells into clusters, starting the process of fusion of the densest cells.

The next cell is included in some cluster if at least one cell neighbor already belongs to the cluster. If the neighbors of the cell do not belong to any formed cluster, then the cell forms a new cluster. If the neighbors of the cell belong to several existing clusters, the respective clusters are merged into a new cluster.

Combining cells into clusters uniquely determines the distribution of multiple points between the clusters. The user must specify a grid step parameter and a minimum grid cell density for which the cluster joining process is not performed. Low-density cells are considered noise.

© A. V. Kuzmin, L. D. Grekov, N. M. Kuzmina, O. A. Petrov, O. M. Medvedenko, 2020

The algorithm does not require a preliminary task of the number of clusters and information about the nature of the distribution of points in the input set.
The proposed algorithm can be used to process large arrays of point data of large spatial resolution. The most promising area of application of the algorithm is the analysis of multispectral satellite images of medium and high resolution in the fields of the analysis of the state of agricultural resources, forest resources and various natural landscapes. The result of clustering the space image data can also be used to create a classifier's training set.
Keywords: *clustering algorithm; satellite space images; grid; cell; point density; neighborhood ratio; agrarian resources; natural landscapes; forest resources*

А.В. Кузьмін[1], Л.Д. Греков[2], Н.М. Кузьміна[3], О.А. Петров[4], О.М. Медведенко[4]

[1]Київський національний університет імені Тараса Шевченка, м. Київ, Україна
[2]ДНВЦ «Природа», м. Київ, Україна
[3]Національний педагогічний університет імені М.П. Драгоманова, м. Київ, Україна
[4]ТОВ «НВП «Агроресурссистеми», м. Київ, Україна

ОБЧИСЛЮВАЛЬНІ ПРОЦЕДУРИ ТЕМАТИЧНОЇ ОБРОБКИ КОСМІЧНИХ ЗНІМКІВ В ІНТЕРЕСАХ МОНІТОРИНГУ АГРАРНИХ РЕСУРСІВ (ЧАСТИНА 2)

Анотація. *Розглядається універсальний швидкий алгоритм кластерного аналізу. Запропонований алгоритм відноситься до сіткового типу, використовує параметр щільності точок в комірці сітки і відношення сусідства для об'єднання сусідніх щільних комірок в кластери. Алгоритм послідовно обчислює для кожної точки номер комірки, якій вона належить, потім формує групи точок для кожної непорожньої комірки. Далі послідовно об'єднує комірки в кластери, починаючи процес об'єднання з найбільш щільних комірок. Чергова комірка включається в деякий кластер, якщо хоча б один сусід комірки вже належить кластеру. Якщо сусіди комірки не належать жодному утвореному кластеру, то комірка утворює новий кластер. У випадку коли сусіди комірки належать зразу декільком існуючим кластерам, відповідні кластери об'єднуються у новий кластер.*
Об'єднання комірок в кластери однозначно визначає розподіл по кластерах множини точок. Для роботи алгоритму користувачу треба задавати параметр кроку сітки та мінімальну щільність комірок сітки, для яких процес приєднання до кластерів не здійснюється. Комірки з малою щільністю вважаються шумом.
Алгоритм не вимагає попереднього завдання кількості кластерів і інформації про характер розподілу точок вхідної множини.
Запропонований алгоритм може використовуватися для обробки великих масивів точкових даних великої просторової розмірності. Найбільш перспективним напрямком застосування алгоритму є аналіз мультиспектральних супутникових знімків середньої та високої розподільчої здатності в інтересах аналізу стану агроресурсів, лісових ресурсів та різноманітних природних ландшафтів. Результат кластеризації даних космічного знімку може також використовуватись для створення навчальної множини класифікатора.
Ключові слова: *алгоритм кластеризації; супутникові космічні знімки; сітка; комірка; щільність точок; відношення сусідства; аграрні ресурси; природні ландшафти; лісові ресурси*

Вступ

Процедури кластеризації відносяться до найбільш важливих завдань інтелектуального аналізу (Data Mining), які припускають розбиття деякої множини точкових елементів на умовно непересічні підмножини – кластери – на основі властивості однорідності і схожості їх характеристик.

Слід зазначити, що в практичних задачах кластеризації кластери часто погано розділяються, мають складну форму, а області їх значень перетинаються, що робить застосування багатьох відомих алгоритмів неефективним.

Особливо важливою сферою застосування методів кластеризації є сегментація мультиспектральних супутникових зображень [1, 2], зокрема для аналізу стану аграрних ресурсів. Серед прикладних задач аналізу аграрних ресурсів, які використовують методи кластерного аналізу, можна виділити:

- Контроль за використанням сільгоспземель. Поділ на кластери: поля, що обробляються, і ті, що не використовуються для вирощування сільськогосподарських культур;
- Контроль за перебігом посівної та жнив. Поділ на кластери: поля, де процес посіву або збирання вже відбувся або ще не почався;
- Поділ полів на кластери озимих та ярих культур;
- Поділ полів озимих культур на зернові та технічні.

У таких прикладних задачах дані характеризуються*:*

- великим обсягом – $10^5 – 10^7$ об’єктів;
- високою просторовою розмірністю даних і різнорідністю їх характеристик;
- відсутністю апріорної інформації про кількість кластерів і ймовірнісні їх характеристики;
- наявністю шуму і викидів у вхідних даних.

Все це призводить до актуальності розробки ефективних методів кластерного аналізу для вирішення прикладних завдань такого типу.

Прийнятний алгоритм кластерного аналізу для задач сегментації супутникових зображень повинен відповідати таким вимогам [3]:

- низька обчислювальна складність;
- можливість виділяти кластери різної структури;
- виділення заздалегідь невідомого числа кластерів;
- можливість обробляти дані при наявності шуму;
- простота настроювання параметрів алгоритму.

У даній роботі розглядається алгоритм кластеризації сіткового типу, який використовує характеристики щільності комірок сітки і принцип сусідства комірок з високою щільністю. Алгоритм не вимагає початкового задавання вихідного числа кластерів і використовує в своїй роботі два пов'язаних між собою параметри, які вибирає користувач:

- h – максимально допустимий крок сітки, який визначає масштаб кластерів, що підлягають виділенню;
- M_0 – щільність комірок, що відсікаються.

Метод, описаний нижче, можна умовно назвати методом об'єднання сусідніх комірок, або *Method of Uniting Neighboring Cells – MUNC (eng.)*

Нехай в n-вимірному евклідовому просторі задана скінченна множина векторів $\Omega = \{\overrightarrow{X}^{(i)} = x_{i,1}, x_{i,2}, ... x_{i,n},\ i = \overline{1,N}\}$, $\overline{X}_{\max} = \{\max\limits_{i=\overline{1,N}} x_{i,j},\ j = \overline{1,n}\}$,

$\overline{X}_{\min} = \{\min\limits_{i=\overline{1,N}} x_{i,j},\ j = \overline{1,n}\}$.

Множина точок Ω належить n-вимірному паралелепіпеду $\Pi = [x_{\min,1}, x_{\max,1}] \times ... [x_{\min,n}, x_{\max,n}]$.

Побудова сітки

Задаємо параметр дискретизації сітки h і обчислимо кількість розбиттів паралелепіпеда Π в напрямку кожної координати:

$m_j = \left[\frac{x_{\max,j} - x_{\min,j}}{h} \right] + 1,\ j = \overline{1,n}$, де [ ] – ціла частина числа. Таким чином, загальна кількість комірок, які покривають паралелепіпед Π, дорівнює $M = \prod_{j=1}^{n} m_j$.

Уточнимо крок сітки по кожній координаті $h_j = \left(x_{\max,j} - x_{\min,j}\right) / m_j,\ j = \overline{1,n}$.

Кожна комірка сітки характеризується своїм цілочисельним векторним індексом розмірності n, $R = \{(r_1, r_2, ... r_n),\ 1 \le r_i \le m_i,\ i = \overline{1,n}\}$. Множина векторних індексів комірок сітки R взаємно-однозначно відображується на множину натуральних чисел: $R \Leftrightarrow Z = \{1,2,...M\}$.

Розподіл точок множини Ω по комірках

Послідовно для кожної точки множини Ω обчислюємо багатовимірний індекс комірки, якій вона належить:

$R^{(i)} = \left\{ r_j^i = \left[\frac{x_{i,j} - x_{\min,j}}{h_j} \right] + 1,\ j = 1,n \right\} \forall i = \overline{1,N}$, а також її одновимірний образ $Z^{(i)}$.

В результаті такого обчислення формується двовимірний список $\Omega_Z = \left\langle \left\langle Z^{(i)}, i \right\rangle, i = \overline{1,N} \right\rangle$, який встановлює приналежність кожної точки множини Ω одній з комірок сітки. Очевидно, що значна кількість комірок сітки при цьому залишаться порожніми, а кількість непорожніх комірок визначається множиною: $Z_F = \left\{ Z^{(i_1)}, Z^{(i_2)}, ... Z^{(i_m)} \right\}$, де кожний елемент

враховується лише один раз, $|Z_F|$ – кількість елементів цієї множини. Зазвичай $|Z_F| << M$.

Далі список Ω_Z сортуємо по ключу першого елемента і формуємо групуючий список наступної структури:

$\Omega_{ZG} = \left\langle Z^{(i)}, \Omega^{(i)}, N^{(i)}, \mu^{(i)}, i = \overline{1, |Z_F|} \right\rangle$, де $\Omega^{(i)}$ – множина номерів точок, які належать комірці з індексом $Z^{(i)}$, $N^{(i)}$ – множина сусідів комірки $Z^{(i)}$, $\mu^{(i)}$ – щільність комірки (кількість точок, що належать комірці).

Список Ω_{ZG} сортуємо за зменшенням ключа $\mu^{(i)}$:

$\Omega_{ZG} = \left\langle Z^{(s_i)}, \Omega^{(s_i)}, N^{(s_i)}, \mu^{(s_i)}, i = \overline{1, |Z_F|} \right\rangle$, де $\mu^{(s_i)} \geq \mu^{(s_{i+1})}$.

Об’єднання комірок в кластери

Об'єднання комірок в кластери однозначно визначає кластеризацію множини точок Ω. Тому достатньо об’єднати в кластери одновимірні індекси комірок $Z^{(s_i)}, i = \overline{1, |Z_F|}$ використовуючи відношення сусідства $N^{(s_i)}$ та значення щільності $\mu^{(s_i)}$.

Індекс найбільш щільної комірки $Z^{(s_1)}$ заноситься до першого кластеру $K_1 = \left\{ Z^{(s_1)} \right\}$.

Обираємо наступну комірку з індексом $Z^{(s_2)}$, перевіряємо виконання умови $K_1 \cap N^{(s_2)} = \otimes$. При виконанні цієї умови, комірка $Z^{(s_2)}$ породжує новий кластер $K_2 = \left\{ Z^{(s_2)} \right\}$, у протилежному випадку $K_1 = K_1 \cup \left\{ Z^{(s_2)} \right\}$, тобто елемент $Z^{(s_2)}$ додається до кластера K_1.

Припустимо, що здійснено m кроків, в результаті яких комірки з індексами $Z^{(s_1)}, Z^{(s_2)}, ..., Z^{(s_m)}$ віднесені до одного з кластерів $K_1, K_2, .. K_r$.

Визначено порядок приналежності комірки з індексом $Z^{(s_{m+1})}$ деякому кластеру.

Обчислимо $K_i \cap N^{(s_{m+1})}, i = \overline{1, r}$ та перевіримо умову $\bigcup_{i=1}^{r} \left(K_i \cap N^{(s_{m+1})} \right) = \otimes$.

При виконанні цієї умови комірка з індексом $Z^{(s_{m+1})}$ утворює новий кластер $K_{r+1} = \left\{ Z^{(s_{m+1})} \right\}$.

При порушенні цієї умови існує $i_1 \leq i_2 \leq ... \leq i_l$ – номери вже утворених кластерів, для яких $K_{i_j} \cap N^{(s_{m+1})} \neq \otimes, j = \overline{1, l}$. В цьому випадку ці кластери

об'єднуються в один кластер K_{i_1} і долучають комірку з номером $Z^{(s_{m+1})}$, тобто $K_{i_1} = K_{i_1} \cup K_{i_2} \cup ... \cup K_{i_l} \cup \left\{Z^{(s_{m+1})}\right\}$. Кластери $K_{i_2}, K_{i_3} ... K_{i_l}$, що увійшли в K_{i_1}, видаляються, а нумерація усіх інших $r - r_l + 1$ кластерів зсувається.

Процес об'єднання комірок в кластери закінчується на кроці m_0, коли виконується умова $\mu^{(m_0)} \le \mathrm{M}_0$. Тобто коли щільність комірок, що залишились, має щільність меншу за деяке порогове значення.

Множину комірок з малою щільністю точок, які не віднесені до жодного кластера, можна залишити некластеризованими, сприймаючи їх як викиди або шум.

Другий варіант завершення процедури кластеризації можна здійснити методом класифікації (навчання з вчителем), де в якості множини навчання використовуються вже сформовані кластери, наприклад методом мінімальної відстані або методом найближчих сусідів.

Тестова перевірка працездатності алгоритму

Тестування алгоритму проводилось на штучно згенерованих точкових множинах у двовимірному просторі.

Множина 1 містить 300 000 точок, які представляють 7 нелінійно розділених кластерів різної форми, різного розміру та щільності. Змодельована кластерна структура була штучно зашумлена множиною з 2000 точок, які покривали увесь простір, що займали кластери (рис. 1).

Процедура кластеризації проводилась для різних значень параметрів налаштування h і M_0. Очікувані і найкращі результати отримані для значень $h \le 0.7$ і $3 \le M_0 \le 10$ (рис. 2). В результаті виконання процедури 299 545 точок були віднесені до кластерів.

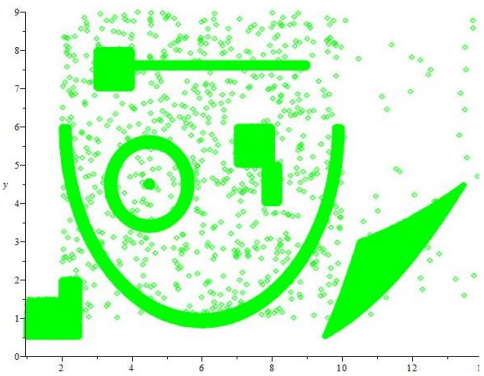

Рис. 1

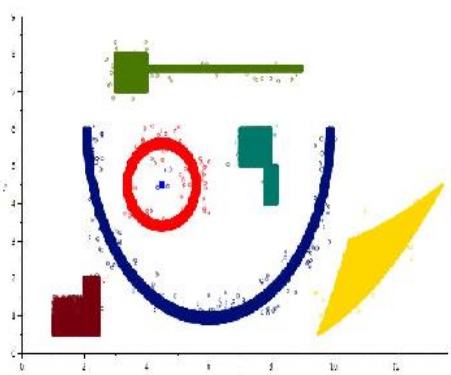

Рис. 2

Множина 2 містить 100 000 точок, об'єднаних у 3 штучно змодельованих кластери еліптичної форми, які мають зони перекриття. Кожний кластер

представляє собою множину точок на площині двовимірного розподілу Гауса з різними векторами середніх, дисперсій і нульовою коваріацією (рис. 3).

Враховуючи, що змодельовані кластери мали зони перекриття, особливо значну між другим и третім (синім та зеленим), алгоритм дозволив чітко виділити ядра кожного кластера, які є найбільш щільними частинами кластерів. Найкращі результати отримані для значень параметрів $h \approx 0.6$ и $M_0 \approx 25$. При цьому до кластерів було віднесено 85% усіх точок множини (рис. 4).

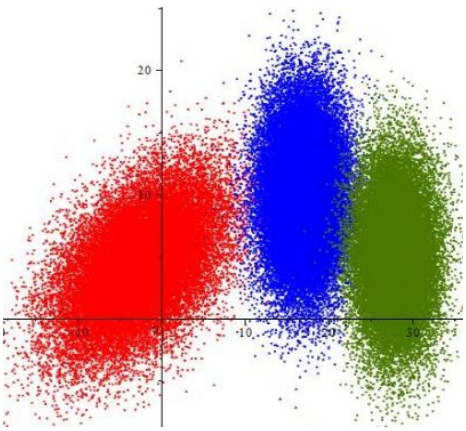

Рис. 3

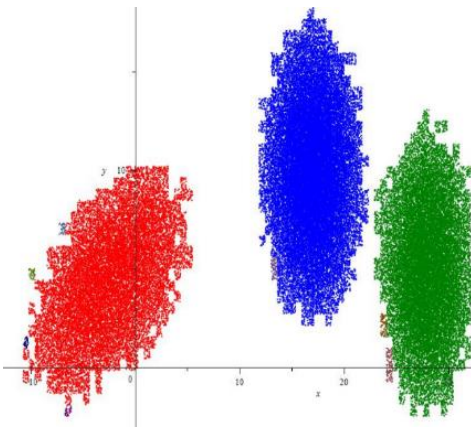

Рис. 4

СПИСОК ЛІТЕРАТУРИ

1. А.В. Кузьмін, Л.Д. Греков, О.А. Петров, О.М. Медведенко (2017) Обчислювальні процедури тематичної обробки космічних знімків в інтересах моніторингу аграрних ресурсів (частина 1). Екологічна безпека та природокористування № 1-2(23), 70-78. https://doi.org/10.32347/2411-4049.2017.1
2. Sarmah S., Bhattacharyya D.K. (2012) A grid-density based technique for finding clusters in satellite image. Pattern Recognition Letters, V. 33, 589-604.
3. И.А. Пестунов, Ю.Н. Синявский (2012) Алгоритм кластеризации в задачах сегментации спутниковых изображений. Вестник Кемеровского государственного университета №4 (52), т. 2, 110-125.

Стаття надійшла до редакції 17.10.2019 і прийнята до друку після рецензування 23.12.2019

REFERENCES

1. Kuzmin, A.V., Hrekov, L.D., Petrov, O.A., & Medvedenko, O.M. (2017). Obchysliuvalni protsedury tematychnoi obrobky kosmichnykh znimkiv v interesakh monitorynhu ahrarnykh resursiv (chastyna 1). *Ekolohichna bezpeka ta pryrodokorystuvannia*, *1-2*(23), 70-78. https://doi.org/10.32347/2411-4049.2017.1
2. Sarmah, S., & Bhattacharyya, D.K. (2012). A grid-density based technique for finding clusters in satellite image. *Pattern Recognition Letters*, 33, 589-604.
3. Pestunov, I.A., & Sinjavskij, Ju.N. (2012). Algoritm klasterizacii v zadachah segmentacii sputnikovyh izobrazhenij. *Vestnik Kemerovskogo gosudarstvennogo universiteta*, *4*(52), t.2, 110-125.

The article was received 17.10.2019 and was accepted after revision 23.12.2019

Кузьмін Анатолій Володимирович
кандидат фізико-математичних наук, доцент факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка
Адреса робоча: Україна, 01601, м. Київ, вул. Володимирська, 64/13
e-mail: kuzmin_a_b@ukr.net
ORCID 0000-0001-5439-6387

Греков Леонід Дмитрович
доктор технічних наук, старший науковий співробітник, директор ДНВЦ «Природа»
Адреса робоча: 03680, Україна, Київ, проспект Акад. Глушкова, буд. 40, корпус 4/1
ORCID 0000-0002-1604-7730

Кузьміна Наталія Миколаївна
кандидат фізико-математичних наук, доцент факультету інформатики Національного педагогічного університету імені М.П. Драгоманова
Адреса робоча: Україна, 01601, м. Київ, вул. Пирогова, 9
e-mail: n.m.kuzmina@npu.edu.ua
ORCID 0000-0003-0136-1441

Петров Олексій Анатолійович
кандидат географічних наук, ТОВ «НВП «Агроресурссистеми»
Адреса робоча: Україна, 01133, м. Київ, пров. Лабораторний, 1, оф. 450
ORCID 0000-0001-9828-2007

Медведенко Олена Миколаївна
директор ТОВ «НВП «Агроресурссистеми»
Адреса робоча: Україна, 01133, м. Київ, пров. Лабораторний, 1, оф. 450
ORCID 0000-0002-9178-9638