УДК 681.3

К.Ю. Новікова, О.І. Михальов

ДОСЛІДЖЕННЯ АЛГОРИТМІВ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ В ЗАДАЧАХ АНАЛІЗУ МЕТАЛОГРАФІЧНИХ ЗОБРАЖЕНЬ

Аннотація. Робота присвячена дослідженню алгоритмів нечіткої кластеризації в задачах аналізу металографічних зображень. Для дослідження в роботі було реалізовано два алгоритми нечіткої кластеризації: с-середніх та субтрактивний метод нечіткої кластеризації для зразків мікроструктур металографічних зображень (x100, x500).

Ключові слова: кластерний аналіз, кластер, нечітка кластеризація, алгоритм fcm, субтрактивна кластеризація, центр кластеру, металографічні зображення.

Вступ. Термін "кластерний аналіз" насправді включає в себе набір різних алгоритмів класифікації. Загальне питання, що ставиться дослідниками у багатьох галузях, полягає в тому, як організувати спостережувані дані в наочні структури, тобто розгорнути таксономії.

Техніка кластеризації застосовується в найрізноманітніших галузях. Хартіган дав прекрасний огляд багатьох опублікованих досліджень, що містять результати, отримані методами кластерного аналізу.

Задачі та цілі кластеризації. Кластеризація (Data clustering) - завдання машинного навчання, в якому потрібно розділити задану вибірку об'єктів (ситуацій) на підмножини, які не перетинаються. Такі підмножини називаються кластерами. Причому ділити слід, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися [1].

Задачі кластеризації відноситься до більш широкого класу завдань навчання без учителя.

Типи вхідних даних:

– ознаковий опис об'єктів. Кожен об'єкт описується набором своїх характеристик, які називаються ознаками. Ознаки можуть бути числовими або нечисловими;

 ISSN 1562-9945

– матриця відстаней між об'єктами. Кожен об'єкт описується відстанями до всієї решти навчальної вибірки.

Цілі кластеризації:

– розуміння даних шляхом виявлення кластерної структури. Розподіл вибірки на групи схожих об'єктів дозволяє спростити подальшу обробку даних і прийняття рішень, застосовуючи до кожного кластеру свій метод аналізу;

– стиснення даних. Якщо вихідна вибірка занадто велика, то можна скоротити її, залишивши по одному найтиповішому представникові від кожного кластера;

– виявлення новизни. Виділяються нетипові об'єкти, які не вдається приєднати до жодного з кластерів.

Постановка задачі нечіткої кластеризації. Концептуальний взаємозв'язок між кластерним аналізом і теорією нечітких множин ґрунтується на тій обставині, що при розв'язанні завдань структуризації складних систем більшість формованих класів об'єктів розмиті за своєю природою.

Один із варіантів конкретизації задачі нечіткого кластерного аналізу, для розв'язання якого може бути використана спеціальна функція fcm системи MATLAB, заснований на алгоритмі її розв'язання методом нечітких с-середніх.

Для уточнення виду цільової функції $f(\Im(A))$ до розгляду вводяться деякі додаткові поняття. Насамперед передбачається, що необхідні нечіткі кластери являють собою нечіткі множини A_k, утворюють нечітке покриття вихідної множини об'єктів кластеризації $A = \mathrm{A}$, набуває такого вигляду:

$$\sum_{k=1}^{c} \mu_{A_k}(a_i) = 1 \quad (\forall a_i \in A), \tag{1}$$

де с - загальна кількість нечітких кластерів $A_k(\kappa \in \{2,\ldots,c\})$, яка вважається попередньо заданою $(c \in \mathbb{N} \ u \ c > 1)$.

Далі для кожного нечіткого кластеру вводяться до розгляду так звані типові представники або центри ν_k потрібних нечітких кластерів $A_k(\kappa \in \{2,\ldots,c\})$, які розраховуються для кожного з нечітких кластерів і за кожною із ознак за такою формулою:

$$v_j^k = \frac{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m \times x_j^i}{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m} \quad (\forall k \in \{2,\ldots,c\},\ \forall p_j \in P), \tag{2}$$

де m - деякий параметр, званий експоненціальною вагою і рівний деякому дійсному числу (m>1). Кожен із центрів кластерів є вектором $v_k = (v_1^k, v_2^k, \ldots, v_q^k)$ в деякому q-мірному нормованому просторі, ізоморфної $\mathbb{R}^q$, *тобто* $v_j^k \in \mathbb{R}^q$, якщо всі ознаки виміряні у шкалі відносин.

Зрештою, в якості цільової функції будемо розглядати суму квадратів зважених відхилень координат об'єктів кластеризації від центрів потрібних нечітких кластерів:

$$f(A_k, v_q^k) = \sum_{i=1}^{n} \sum_{k=1}^{c} (\mu_{A_k}(a_i))^m \sum_{j=1}^{q} (x_j^i - v_j^k)^2, \tag{3}$$

де m - експонентна вага нечіткої кластеризації $m \in \mathbb{R}, m > 1$, значення якої задається в залежно від кількості елементів (потужності) множини A. Чим більше елементів містить множина A, тим менше значення вибирається для m.

Завдання нечіткої кластеризації. Завдання нечіткої кластеризації може бути сформульовано таким чином: для заданих матриці даних D, кількості нечітких кластерів c $c \in \mathbb{N}, c > 1$, параметра m визначити матрицю U значень функцій належності об'єктів кластеризації $a_i \in A$, нечітким кластерам A_k $(k \in \{2,\ldots,c\})$, які доставляють мінімум цільової функції (3) і задовольняють обмеженням (1), (2), а також додатковим обмеженням (4) і (5):

$$\sum_{i=1}^{n} \mu_{A_k}(a_i) > 0 \quad (\forall\ k \in \{2,\ldots,c\}), \tag{4}$$

$$\mu_{A_k}(a_i) \geq 0 \quad (\forall k \in \{2,\ldots,c\},\ \forall a_i \in A). \tag{5}$$

Умова (4) виключає появу порожніх нечітких кластерів в потрібній нечіткій кластеризації. Остання умова (5) має суто формальний характер, оскільки безпосередньо випливає з визначення функції належності нечітких множин. У цьому випадку мінімізація цільової функції (3) мінімізує відхилення всіх об'єктів кластеризації від центрів нечітких кластерів пропорційно значенням функцій належності цих об'єктів до відповідних нечітких кластерів.

Оскільки цільова функція (2) не є опуклою, а обмеження (1), (2), (4), (5) у своїй сукупності формують неопуклі множини припустимих альтернатив, то в загальному випадку завдання нечіткої кластеризації відноситься до багатоекстремальних завдань нелінійного програмування.

Перевагою постановки задачі нечіткої кластеризації у вигляді (1) - (5) є природна інтерпретація як потрібних нечітких кластерів, що визначаються функціями належності (5), так і їхніх типових представників або центрів (2), які також визначаються в результаті розв'язання поставленої задачі.

Недоліком такої постановки задачі нечіткої кластеризації є необхідність апріорного завдання загального числа нечітких кластерів с $c \in \mathbb{N}, c > 1$, яке в окремих випадках може бути невідомо. Ця обставина може вимагати залучення додаткових процедур для її визначення або розв'язання поставленої задачі для декількох значень с з подальшим вибором найбільш адекватного результату нечіткої кластеризації [2].

Постановка задачі. Дослідити рівні експертної інформації щодо металографічних зображень колісної сталі.

Дослідити роботу алгоритмів нечіткої кластеризації в задачах аналізу металографічних зображень.

Дослідження алгоритмів нечіткої кластеризації в задачах аналізу металографічних зображень. Для аналізу металографічних зображень були обрані три зразки зображень колісної сталі:

– ліквацийна неоднорідність мікроструктури, зразок № 111 (плавка №42380), полоса мартенсіта;

– ліквацийна неоднорідність мікроструктури, х100, зразок № 12 (плавка № 22325), полоса мартенсіта;

– мікроструктура колесної сталі, х500, зразок №102 (плавка № 21384).

Розмір даних зображень складає 640х480. Для розв'язання задачі нечіткої кластеризації реалізовані два алгоритми:

– алгоритм FCM;

– алгоритм субтрактивної нечіткої кластеризації.

На рисунку 1 представлено перший зразок мікроструктури колісної сталі.

Рисунок 1 – Зразок №111 (пл. №42380), смуга мартенсіта

Після отримання першого зображення ми створюємо цикл за допомогою якого отримуємо зображення з поданням про значення кожного пікселя. Отриманий в результаті файл будемо використовувати для подальшого дослідження, тому що дані для кластеризації повинні бути представлені в вигляді матриці чисел.

Наступним кроком дослідження є визначення центрів кластеризації, для цього застосовується функція fcm, яка виконується ітераційно до тих пір, поки зміни цільової функції перевищують деякий заданий поріг. На кожному кроці в командному вікні виводяться порядковий номер ітерації і відповідне поточне значення цільової функції. Якщо після запису функції fcm у другому рядку не ставити крапку з комою (;), то у вікні команд будуть показані значення координат центрів нечітких кластерів, значення функцій приналежності об'єктів нечітким кластерам і значення цільової функції на кожній з ітерацій роботи алгоритму FCM. У цьому прикладі використовується перший формат запису функції fcm.

```
>> [center,U,obj_fcn]=fcm(dip,4);
Iteration count = 1, obj. fcn = 511325634.087258
Iteration count = 2, obj. fcn = 397350961.226726
Iteration count = 3, obj. fcn = 397324432.614382
Iteration count = 4, obj. fcn = 397324264.686743
Iteration count = 5, obj. fcn = 397324261.917665
Iteration count = 6, obj. fcn = 397324261.856672
Iteration count = 7, obj. fcn = 397324261.855243
Iteration count = 8, obj. fcn = 397324261.855209
Iteration count = 9, obj. fcn = 397324261.855208
>> plot(obj_fcn);
```

Для оцінки динаміки зміни значень цільової функції використовуємо команду побудови графіка plot(obj_fcn). Результати

показані на рисунку 2, з рисунку видно, що чим вище значення цільової функції fcn тим краще кількість ітерацій нечіткої кластеризації.

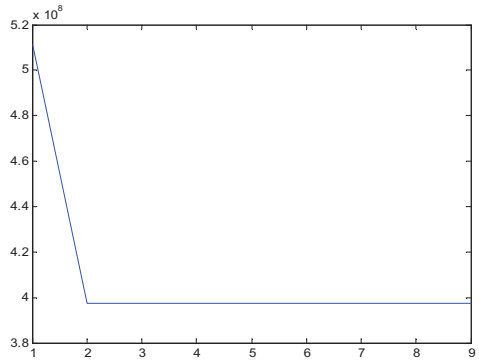
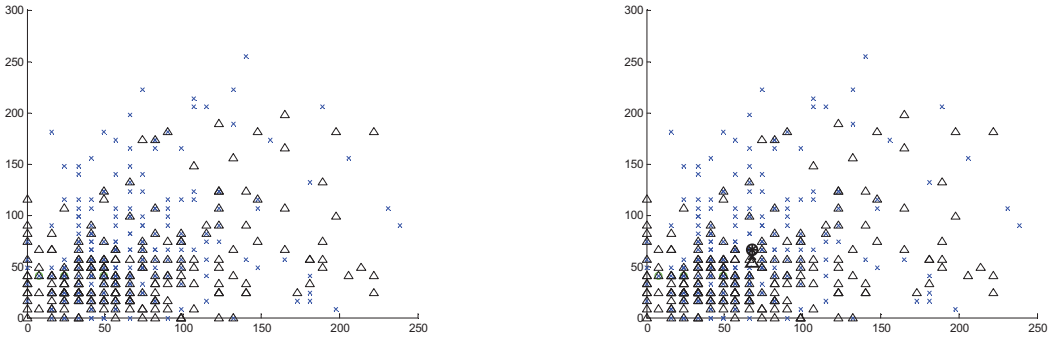


Рисунок 2 – Графік зміни значень цільової функції

Далі визначаємо максимальну степінь приналежності окремого елемента даних кластеру.

Результат розв'язання задачі нечіткої кластеризації для 4-х нечітких кластерів із використанням зазначеної послідовності команд може бути візуалізовано на рисунку 3, де представлено множини даних, що підлягають кластеризації з знайденими центрами кластерів.



а) б)

а) множина аналізованих даних

б) множина аналізованих даних та центри кластерів

Рисунок 3 – Результати роботи алгоритму нечіткої кластеризації fcm

На рисунку 4 відображені результати роботи нечіткої кластеризації з розбивкой на нечіткі кластера (функція приналежності – розбивка по яскравості, вид - трапецевидні). Візуалізація спрощує задачу експерта з аналізу зображень та зменшує ймовірність помилки.

Результати роботи нечіткої кластеризації

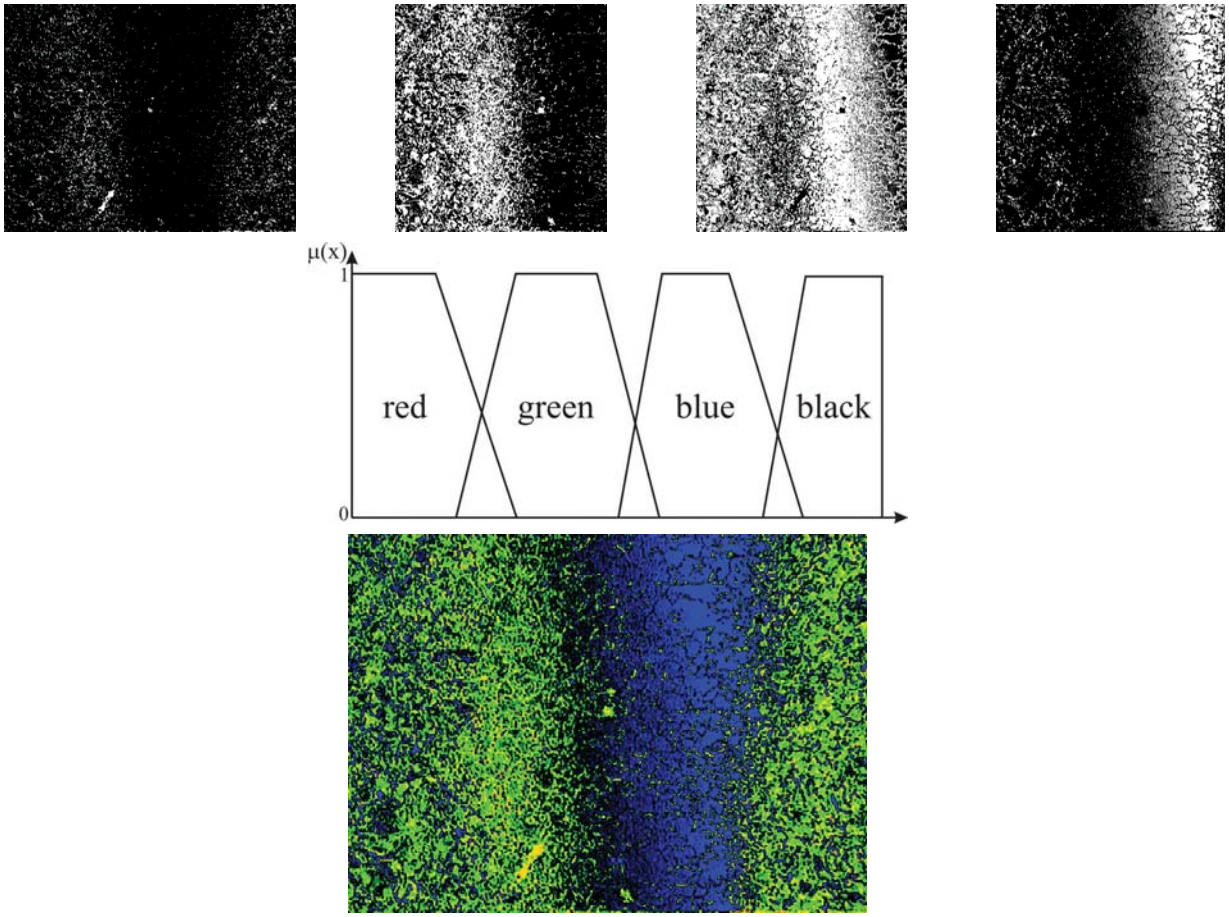


Рисунок 4

Проведемо дослідження алгоритму fcm для аналізу чорно – білого зображення зразка №111 (пл. №42380), смуга мартенсіта.

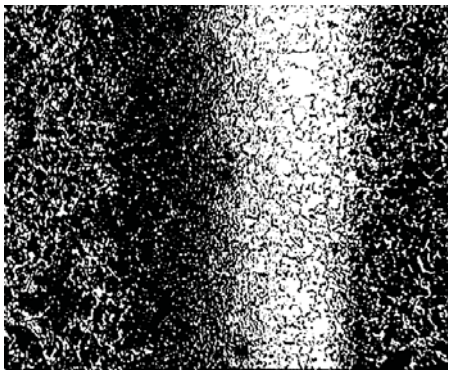


Рисунок 5 – Чорно-біле зображення зразку №111 (пл. №42380), смуга мартенсіта

Після отримання чорного-білого зображення ми створюємо цикл за допомогою якого отримуємо зображення з поданням про значення кожного пікселя. Отриманий в результаті файл будемо використовувати для подальшого дослідження, тому що дані для кластеризації повинні бути представлені в вигляді матриці чисел.

 ISSN 1562-9945

Наступним кроком дослідження є визначення центрів кластеризації, для цього застосовується функція fcm, яка виконується ітераційно до тих пір, поки зміни цільової функції перевищують деякий заданий поріг. На кожному кроці в командному вікні Matlab виводяться порядковий номер ітерації і відповідне поточне значення цільової функції. Якщо після запису функції fcm у другому рядку не ставити крапку з комою (;), то у вікні команд будуть показані значення координат центрів нечітких кластерів, значення функцій приналежності об'єктів нечітким кластерам і значення цільової функції на кожній з ітерацій роботи алгоритму FCM. У цьому прикладі використовується перший формат запису функції fcm.

```
>> [center,U,obj_fcn]=fcm(diplom,2);
Iteration count = 1, obj. fcn = 29667.404451
Iteration count = 2, obj. fcn = 24601.732554
Iteration count = 3, obj. fcn = 24600.740623
Iteration count = 4, obj. fcn = 24600.735477
Iteration count = 5, obj. fcn = 24600.735418
Iteration count = 6, obj. fcn = 24600.735417
>> plot(obj_fcn);
```

Для оцінки динаміки зміни значень цільової функції використовується команда побудови графіка plot(obj_fcn). Результати показані на рисунку 6, з рисунку видно, що чим вище значення цільової функції fcn тим краще кількість ітерацій нечіткої кластеризації.

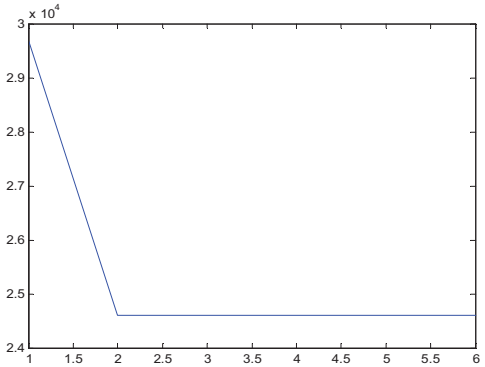


Рисунок 6 – Графік зміни значень цільової функції

Далі визначаємо максимальну степінь приналежності окремого елемента даних кластеру.

Результат розв'язання задачі нечіткої кластеризації для 2-х нечітких кластерів із використанням зазначеної послідовності команд

може бути візуалізовано на рисунку 7, де представлено множини даних, що підлягають кластеризації з знайденими центрами кластерів.

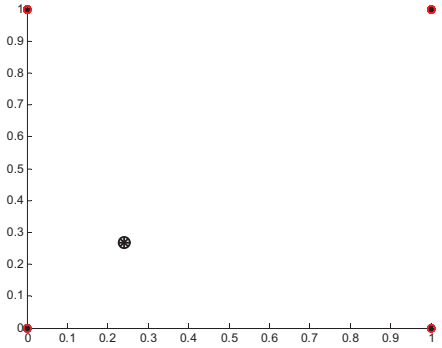


Рисунок 7 – Результат розв'язання задачі нечіткої кластеризації для 2-х нечітких кластерів

В процесі виконання дослідження було виявлено, що для аналізу кольорових зображень неможливо застосувати алгоритм субтрактивної кластеризації, тому що вхідна матриця даних повинна мати розмір nx2, далі дослідження проводились вже для чорно-білого зображення.

Для того, щоб зображення представити в потрібному форматі для подальшого дослідження, використовуючи алгоритм субтрактивної кластеризації було написано M-файл convert.m, який присвоює чорному кольору – нуль, а білому – одиницю.

У разі відсутності будь-яких апріорних припущень щодо кількості нечітких кластерів можна використовувати функцію командного рядка subclust. В результаті виконання цього фрагмента команд будуть одержані значення матриці центрів кластерів і вектора σ - значень.

```
>> [c,s]=subclust(im,[0.5 0.5],[],[1.25 0.5 0.15 1])
Normalizing data...
Computing potential for each data point...
Found cluster 1, potential = 1
Found cluster 2, potential = 0.600319
Found cluster 3, potential = 0.562503
Found cluster 4, potential = 0.457642
Found cluster 5, potential = 0.395276
c =
  256.0000  189.0000
  315.0000  543.0000
   68.0000  172.0000
  433.0000  174.0000
```

 ISSN 1562-9945

```
  100.0000  559.0000
s = 84.6760  112.9603
```

Як можна помітити, для наведених значень аргументів розглянута функція subcluster знаходить п’ять нечітких кластерів та відображає координати їхніх центрів у командному вікні системи [3].

Висновки. Система MATLAB дозволяє вирішувати задачі нечіткої кластеризації двома способами: алгоритм FCM; алгоритм субтрактивної нечіткої кластеризації. Результати нечіткої кластеризації мають наближений характер і можуть служити лише для попередньої структуризації інформації, що міститься в множині вихідних даних. Вирішуючи задачі нечіткої кластеризації, потрібно пам'ятати про особливості та обмеження процесу вимірювання ознак у сукупності об'єктів кластеризації.

Оскільки нечіткі кластери формуються на основі евклідової метрики, відповідний простір ознак має задовольняти аксіомам метричного простору. У той же час для пошуку закономірностей в проблемній області, що мають не метричний характер, необхідно використовувати спеціальні засоби та інструментарій, розроблені для інтелектуального аналізу даних (Data Mining).

Алгоритми нечіткої кластеризації дозволяють розбити досліджувані об'єкти на нечіткі кластери, а нечіткі кластери в свою чергу можуть бути візуалізовані шляхом управління рівнем яскравості відповідних кольорів, аналогічним функції приналежності. Візуалізація спрощує завдання експерта з аналізу зображень і зменшує ймовірність помилки.

ЛІТЕРАТУРА

1. Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980. – 392 с.
2. Мандель Н.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
3. Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH. – СПб.: БХВ – Петербург, 2003. – 736с.: ил.