УДК 004.93’1(045)

КЛАСИФІКАЦІЯ ОБ’ЄКТІВ НА ПІДСТАВІ ТЕНЗОРНИХ МОДЕЛЕЙ

Ю. М. Мінаєв, Д. М. Вінник , А. В. Гайдай, Д. В. Апонасенко, Є. О. Гончарова

Національний авіаційний університет

minaev@rambler.ru

Розглянуто питання застосування методів нечіткої та ієрархічної кластеризації для ідентифікації об’єктів, моделі яких представлені у вигляді інваріантів бінарних тензорів. У середовищі MATLAB розроблені програмні засоби, що дають змогу виконати класифікацію реальних об’єктів (літаків).

The questions of using the methods fuzzy and hierarchical clusterizations for object identifications, which models are present in the manner of invariants of binary tensors. Software programs in the medium of MatLab, allowing execute a categorization of real objects (planes), is designed.

Вступ

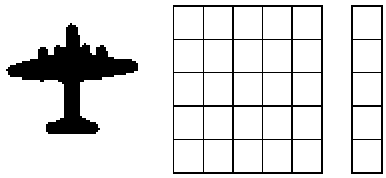
Проблема розпізнавання образів (класифікація об’єктів) належить до класу важкоформалізованих. Як показують дослідження її розв’язання може бути досягнуто на шляху застосування інтелектуальних методів і технологій. Нині існує досить багато методів та алгоритмів розпізнавання образів, але складність проблеми потребує пошуку нових. Головним при цьому є визначення ознак, за якими сформовано алгоритми, та врахування невизначеності, яка притаманна об’єктам, які намагаються класифікувати або розпізнавати.

Сучасний рівень досліджень

Застосування інтелектуальних технологій (ІТ) для розв’язання задач класифікації нині позв’язують з застосуванням обчислювального інтелекту, який також пов’язують з теорією нечітких множин та нейрокомп’ютингом. Одним з ефективних методів класифікації об’єктів є кластеризація, яка, зокрема, вміщує ієрархічні та нечіткі методи. Наявність кластерів, які найчастіше представлено у вигляді дерева, дає змогу сформулювати систему правил, на підставі якої з наявної множини даних беруть знання, тобто певні приховані закономірності, які притаманні об’єкту. Кожний з цих методів має свої недоліки та переваги, алгоритми досить часто мають евристичний характер, але, як показано в працях [1—3], у багатьох випадках задачу класифікації розв’язати вдається. Надзвичайно важливим є визначення системи ознак, які до того ж можуть бути загамованими, або визначеними нечітко (у вигляді інтервалів) і т. ін.

Тензорний підхід до визначення ознак об’єктів для їх класифікації

Для класифікації системи об’єктів застосовують представлення його у вигляді бінарної матриці (яка є проекцією тензора), яку потім розкладають за стовпцями або рядками, які у надалі використовують як вхідні дані. У роботі запропоновано представлення об’єкта тензором, для якого визначають множину інваріантів, які виступають у ролі вхідних ознак. Такий підхід зумовлений тим, що у тензорної моделі об’єкта при зміні координат компоненти тензора змінюються, хоча властивості тензора залишаються інваріантними. Властивості тензора, що залишаються незмінними при перетвореннях координат, визначаються системою його *інваріантів*, що являють собою константи, значення яких зберігаються при зміні системи координат. Величини головних інваріантів можуть бути визначені через власні значення. На рис. 1 наведено приклад представлення об’єкта у вигляді бінарної згорнутої матриці та подано її власні значення.



1 *2* *3*

Рис. 1. Тензорна модель об’єкта:
1 — об’єкт; *2* — тензорна модель;
3 — інваріанти тензорної моделі

Тензорне моделювання об’єктів [4] дає змогу істотно розширити множину визначних властивостей, яким характеризується об’єкт, тому що тензор допускає розкладання на групу так званих «приєднаних тензорів», що дає можливість додаткової інформованості про об’єкт. Властивості тензора — наявність інваріантів, що однозначно характеризують його, можливість згортки тензора — дозволять підвищити конструктивність алгоритмів класифікації.

Кластеризація на основі нечіткого *c*-means алгоритму

Задача нечіткої кластеризації ставиться в такий спосіб [5]. Дано: $X = (X_1, X_2, \ldots, X_n)^T$ — об’єкти, що підлягають кластеризації (n — кількість об’єктів).

Кожний об’єкт $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$ являє собою крапку в p-вимірному просторі ознак (); c — кількість кластерів $2 \le c < n$. Кожному елементу множини X необхідно поставити у відповідність ступені належності c класам.

Елементи одного кластера повинні бути так максимально близькими один до другого в той час, як одночасно кластери мають бути на найбільшому віддаленні один від одного. Для забезпечення керованості процесу кластеризації необхідно використовувати міру близькості. Як таку зазвичай визначають відстань між двома об'єктами (крапками в *p*-вимірному просторі) X_k і X_l у вигляді речовинної функції d: $X \times X \to R^+$, таку, за якої виконуються умови:

$$d(X_k, X_l) = d_{kl} \geq 0;$$

$$d_{kl} = 0 \Leftrightarrow X_k = X_l;$$

$$d_{kl} = d_{lk}.$$

Додатково, якщо функція d задовольняє правила трикутника, тобто $d_{kl} \leq d_{kj} + d_{jl}$, тоді ця функція є метрикою, хоча виконання цієї властивості не завжди необхідно для задач кластеризації. Розбиття множини $X = (X_1, X_2, \ldots, X_n)^T$ на нечіткі підмножини S_i ($i = \overline{1,c}$), як відомо, повністю визначається функцією належності $\mu_{Si}: X \to [0,1]$.

Припустимо, що μ_{ik} — ступінь належності об'єкта $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$ до підмножини S_i, тобто $\mu_{ik} = \mu_{S_i}(X_k)$, і M — множина всіх дійсних матриць розмірністю $c \times n$. Згідно з працею [1] нечіткою *c*-розбивкою (або матрицею ступенів належності) називається матриця $M = [\mu_{ik}] \in V_{cn}$, для якої виконуються умови:

$$\mu_{ik} \in [0,1], i = \overline{1,c}, k = \overline{1,n};$$

$$\sum_{k=1}^{n} \mu_{ik} = 1,\ k = 1,n;$$

$$\sum_{k=1}^{n} \mu_{ik} \in (0,n),\ i = \overline{1,c}.$$

Зазначимо, що на відміну від чіткої, у разі нечіткої *c*-розбивки будь-який об'єкт одночасно належить до різних (практично всіх) кластерів, але з різним ступенем. Вищенаведені умови потребують, щоб сума ступенів належності об'єкта до всіх кластерів була нормалізована до 1, а також, щоб кількість кластерів, до яких належить об'єкт, не перевищувала заданої *c*.

Припустимо, центри кластерів, тобто крапки в *p*-вимірному просторі, навколо яких сконцентровані відповідні об'єкти, визначені і позначимо їх через $V_i = (V_{i1}, V_{i2}, \ldots, V_{ip})$, $i = \overline{1,c}$. У разі використання евклідової відстані завдання нечіткої кластеризації полягає у визначенні такої матриці ступенів належності M і таких координат центрів кластерів $V = (V_1, V_2, \ldots, V_c)$, що забезпечують мінімум такого критерію:

$$\sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^{m} \left\| X_k - V_i \right\|^2 \to \min,$$

де $V_i = \dfrac{1}{\sum_{k=1}^{n} \mu_{ik}} \sum_{k=1}^{n} (\mu_{ik})^m x_k$ — центр *i*-го кластера, $i = \overline{1,c}$; m — так звана експонентна вага ($m \geq 1$).

Значення експонентної ваги встановлюється до початку кластеризації. Експонентна вага m впливає на матрицю ступенів належності. Чим більше m, тим кінцева матриця *c*-розбивки стає більш «розмитою», і при $m \to \infty$ вона набуває вигляду $M = \left[\dfrac{1}{c}\right]$, що є незадовільним з погляду прийняття рішення, бо всі об'єкти належать до всіх кластерів з одним і тим самим ступенем. Також експонентна вага дає змогу при формуванні координат центрів кластерів підсилити вплив об'єктів з великими значеннями ступенів належності і зменшити вплив об'єктів з малими значеннями ступенів належності. Сучасний рівень досліджень не дає теоретично обґрунтованого правила вибору значення m, звичайно беруть $m = 2$.

Крім того, слід зазначити, що аналітичного розв'язання задачі визначення оптимальних координат центрів кластерів і матриці ступенів належності не існує, тому вона розв'язується чисельно. Один з ітераційних алгоритмів розв'язання цієї задачі реалізований у функції MATLAB fcm, яка стандартно викликається у вигляді — [V, M, obj_fcn] = fcm(X, c, options):

вхідні аргументи:

X — матриця, що представляє дані, які підлягають кластеризації, кожний рядок матриці відповідає одному об'єкту (образу);

c — кількість кластерів, що має бути отримана в результаті виконання функції fcm, кількість кластерів повинно бути більше 1 і менше кількості образів, заданих матрицею X;

options — необов'язковий аргумент, що встановлює параметри алгоритму кластеризації: за умовчанням — 1.

Алгоритм кластеризації зупиняється, коли виконана максимальна кількість ітерацій, або коли поліпшення значення цільової функції за одну ітерацію менше зазначеного мінімально допустимого значення.

Функція fcm має три вихідні аргументи:

V — матриця координат центрів кластерів, отриманих у результаті кластеризації. Кожен рядок матриці відповідає центру одного кластера;

M — матриця ступенів належності образів до кластерів. Кожен рядок матриці відповідає функції належності одного кластера.

obj_fcn — вектор значень цільової функції на кожній ітерації алгоритму кластеризації.

Синтез нечіткої бази знань на основі кластеризації

Позначимо через $V_1, V_2, \ldots, V_c$ — центри кластерів, знайдені в результаті кластеризації. Для спрощення викладень вважатимемо, що центри кластерів задано двома координатами: $V_i = (x_i, y_i)$, $i = \overline{1,c}$. Завдання полягає в синтезі нечітких правил, що зв'язують вхід (x) з виходом (y).

Згідно з парадигмою центру кластера ($V_i(i=\overline{1,c})$) ставиться у відповідність одне нечітке правило типу: якщо $x=\tilde{x}$, то $y=\tilde{y}$, причому нечіткі терми інтерпретуються у такий спосіб: $\tilde{x}_i$ — околиця x_i, $\tilde{y}_i$ — околиця y_i. Функції належностей цих нечітких термів задаються гауссівською кривою:

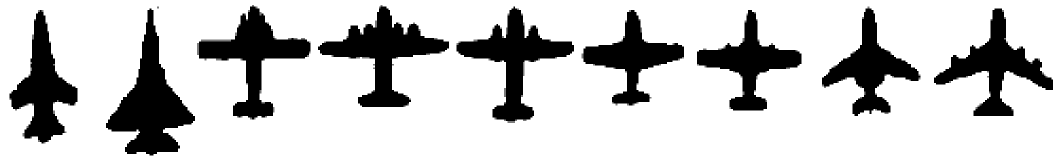
$$\tilde{x}=\exp\left(-\frac{1}{2}\left(\frac{x-x_i}{\beta}\right)^2\right),$$

$$\tilde{y}=\exp\left(-\frac{1}{2}\left(\frac{y-y_i}{\beta}\right)^2\right),\ i=\overline{1,c}$$

де β — параметр алгоритму кластеризації.

Прикладні результати класифікації системи літаків

Групу об’єктів для класифікації показано на рис. 2. Для спрощення аналізу об’єкти представлені у вигляді чотирьох підгруп.



1 *2* *3* *4*

Рис. 2. Множина об’єктів та їх класифікаційні групи

На рис. 3 показана схема кластеризації

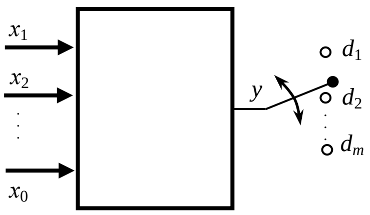


Рис. 3. Схема кластеризації:
$x_1, \dots, x_{1n}$ — сукупність ознак (інваріанти);
$d_1, \dots, d_4$ — групи кластерів

Кожне зображення образу мало вимірність 64x64, кожний образ є вектором власних значень, представлених у 64-вимірному просторі. На рис. 4 наведено результат кластеризації — центри мас кластерів, що візуально виглядають як усереднені образи.

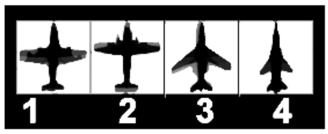


Рис. 4. Кластери, отримані за алгоритмом *fuzzy c-means*

На рис. 5 подано ступені належності до цих кластерів кількох образів з вхідного набору даних.

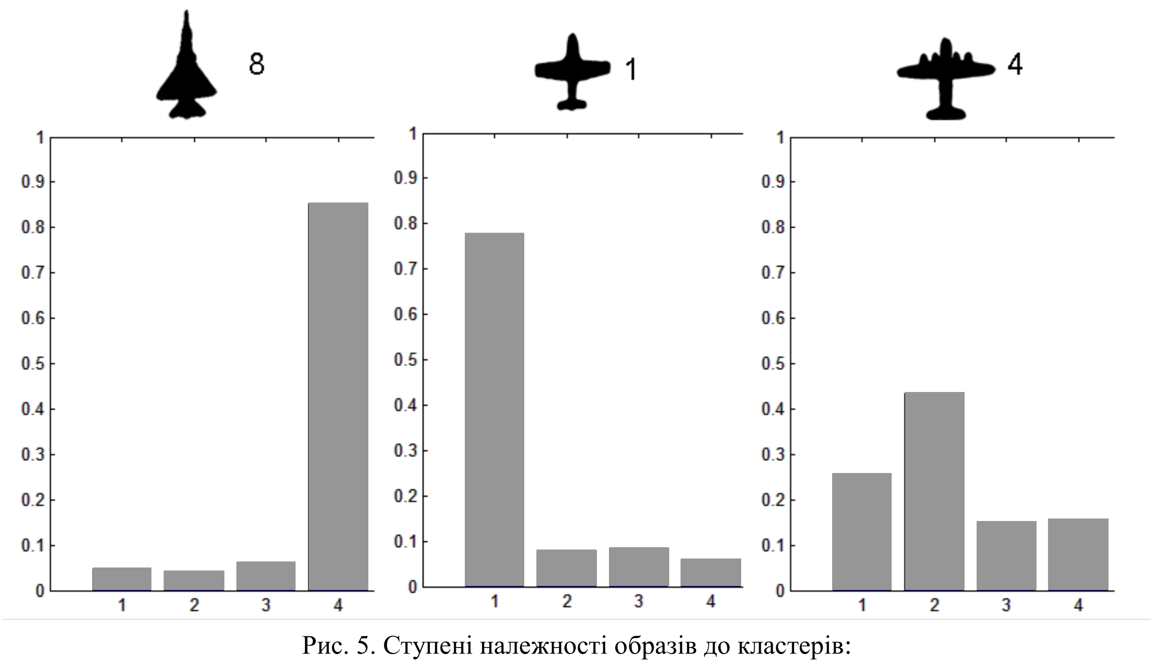
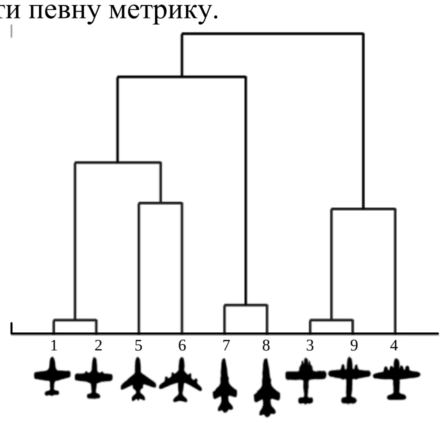


Рис. 5. Ступені належності образів до кластерів:
1 — образ цілком належить кластеру; 0 — цілком не належить

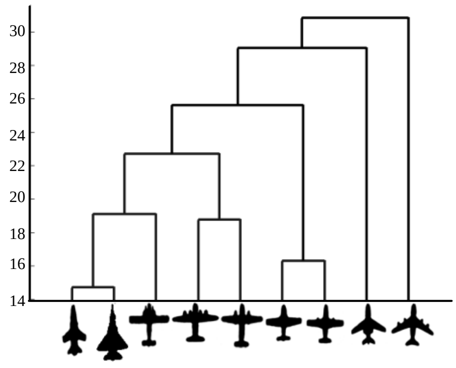
Ієрархічна кластеризація

Дендрограма (бінарне дерево), що використовується для представлення результатів ієрархічної кластеризації, показує ступінь близькості окремих об’єктів і кластерів, а також представляє в графічному вигляді послідовність їх об’єднання або розділення (функція MATLAB — DENDROGRAM(Z), де Z — (M-1) × 3 — матриця, яка представляє кількість спостережень даних і формується за допомогою функції LINKAGE(Y, *method*), яка обчислює ієрархічну кластерну інформацію, використовуючи спеціальні методи визначення відстані).

На рис. 6 подано дендрограми для множини об’єктів. На горизонтальній осі зображено об’єкти, над якими необхідно провести кластеризацію. На вертикальній осі позначено відстань між елементами. Для розрахунку цієї відстані необхідно вибрати певну метрику.



а



б

Рис. 6. Ієрархічна кластеризація:
а — дендрограма тензорно-бінарних зображень об’єктів;
б — дендрограма векторів власних значень тензорно-бінарних зображень об’єктів

У цьому дослідженні використовується евклідова метрика. Чим довше лінії на дендрограмі, що відповідають об’єктам, тим більша відстань між ними і тим більше ці об’єкти розрізняються. Отриману дендрограму можна використати для поділу початкової множини об’єктів на певну кількість кластерів. Це робиться шляхом відсікання частини дендрограми, що нижча деякого порогового значення міжкластерної відстані.

Для того, щоб зменшити кількість обчислень при кластеризації, кожний образ (тензор) представлено вектором власних значень тобто в даному випадку координатами точки 64-вимірного простору. Слід зазначити, що не існує єдиного абсолютно точного алгоритму варіанта кластеризації — результат буде залежати від обраної метрики, алгоритму кластеризації та ін. Критерієм правильності можуть служити близькі результати, що отримані різними методами.

Висновки. 1. Одним з ефективних методів класифікації об’єктів є кластеризація, яка, зокрема, вміщує ієрархічні та нечіткі методи. Наявність кластерів, які найбільш часто представлені у вигляді дерева, дає змогу сформулювати систему правил, на підставі котрої з наявної множини даних беруть знання, що дає можливість виявити приховані закономірності, які притаманні об’єкту.

2. У роботі запропоновано представлення об’єкта класифікації тензором, для якого визначають множину інваріантів, які виступають у ролі вхідних ознак. Властивості тензора, що залишаються незмінними при перетвореннях координат, визначаються системою його *інваріантів*, що являють собою константи, значення яких зберігаються при зміні системи координат.

3. Тензорне моделювання об’єктів дає змогу істотно розширити множини властивостей, якими характеризується об’єкт, за рахунок розкладання головного тензорана, групу «приєднаних тензорів», що дає можливість додаткової інформованості про об’єкт.

4. Властивості тензора — наявність інваріантів, що однозначно характеризують його, можливість згортки тензора — дозволяють підвищити конструктивність алгоритмів класифікації. Проведені експериментальні дослідження класифікації об’єктів у середовищі MATLAB підтвердили ефективність алгоритмів.

ЛІТЕРАТУРА

1. *Xei X. L.* Validity Measure for Fuzzy Clustering / X. L. Xei, G. A. Beni // IEEE Trans. on Pattern Anal. and Machine Intell. 3 (8). — 1991. — P. 841—846.

2. *Yager R.* Essentials of Fuzzy Modeling and Control / R. Yager, D. Filev // USA: John Wiley & Sons. — 1984. — 387 p.

3. *Заде Л. А.* Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. А. Заде. — М. : Мир, 1976. — 165 с.

4. *Минаев Ю. Н.* Тензорный базис как основа новых алгоритмов решения задач управления в условиях неопределенности / Ю. Н. Минаев, О. Ю. Филимонова. — В кн. Новые информационные технологии: Сб. трудов VI Всероссийской науч.-техн. конф. (Москва, 23—24 апреля 2003 г.). В 2-х кн.Т.1/ Под общ. ред. А.П.Хныкина.— М.: МГАПИ, 2003. — С. 142—147.

5. *Zimmermann H.-J.* Fuzzy Set Theory — and Its Applications.3rd ed. / H.-J. Zimmermann. — Kluwer Academic Publishers, 1996. — 435 p.

Стаття надійшла до редакції 5.10.09.