

Іншими перевагами цього обчислення є можливість розпаралелювання та гнучкість в послідовності виконання циклічних згорток. Запропонований підхід актуальний в своєму подальшому розвитку для розподілених паралельних систем з виконанням мікрооперацій великозернистого паралелізму.

1. *Оппенгейм Э. Применение цифровой обработки сигналов.* – М.: Мир, 1980. 2. *Процько І.О. Підхід ефективного обчислення дискретних гармонічних перетворень через циклічні згортки // Вісник НУ "ЛП" "Комп'ютерні системи проектування. Теорія і практика". – 2008. – №626.* 3. *Prots'ko I., The generalized technique of computation the discrete harmonic transforms // Proceeding of the IVth International Conference of Young Scientists MEMSTECH 2008, Poljana, may, 2008.* 4. *Процько І.О. Ефективне обчислення дискретних косинусних перетворень / Вісник НУ "ЛП" "Комп'ютерні системи проектування. Теорія і практика". – 2007. – №591.* 5. *Процько І.О. Підхід ефективного обчислення дискретного перетворення Хартлі / Вісник НУ "ЛП" "Комп'ютерні системи проектування. Теорія і практика". – 2005. – №502.* 6. *Нуссбаумер Г. Быстрое преобразование Фурье и алгоритмы вычисления сверток.* – М.: Радио и связь, 1985. 7. *Макклеллан Дж.Х., Рейдер Ч.М. Применение теории чисел в цифровой обработке сигналов: пер. с англ.* – М.: Радио и связь, 1983.

УДК 681.322

Ю.В. Стех, Файсал М.Е. Сардіх, А.Б. Керницький, М.С. Домброва
Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування

АЛГОРИТМІЧНА ОЦІНКА ОПТИМАЛЬНОСТІ РЕЗУЛЬТАТІВ КЛАСТЕРИЗАЦІЇ ЗА КРИТЕРІЄМ ВІДСТАНІ

© *Стех Ю.В., Файсал М.Е. Сардіх, Керницький А.Б., Домброва М.С., 2010*

Запропоновано алгоритмічний підхід до оцінки оптимальності результатів кластеризації за критерієм відстані. Підхід ґрунтується на розв'язанні двокритеріальної оптимізації за допомогою побудови та оцінювання критеріїв оптимальності у вигляді індексних функцій.

Ключові слова – кластер, образ, індексна функція.

The algorithmic approach to the evaluation optimality of clustering results by distance criterion is proposed. The approach is based on solving two-criterion optimization by building and evaluating the optimality criteria in the form of index functions.

Keywords – cluster, pattern, index function.

Вступ

Кластеризація – об'єднання в групи подібних об'єктів – є однією із фундаментальних задач в області аналізу даних, Data Mining, Web Mining, Text Mining, машинного навчання. Кластеризація часто виступає першим кроком під час аналізу даних. Після виділення кластерів застосовуються інші методи, для кожного кластера будується окрема модель [1; 2; 4].

Задача кластеризації полягає у розбитті загальної множини образів на кластери-класи так, щоб кожен кластер складався з подібних об'єктів, а об'єкти різних кластерів істотно відрізнялися. Сьогодні існує кілька десятків алгоритмів кластеризації і ще більше їх модифікацій.

Результат роботи кожного із алгоритмів кластеризації залежить від багатьох початкових параметрів. Тому знаходження оптимального розбиття множини образів на кластери вимагає побудови і дослідження критеріальних функцій, які дають змогу оцінити результати процесу кластеризації.

Формальна постановка задачі

Нехай $\mathbf{P} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ – множина n образів, кожен з яких має d ознаки. Задача розбиття множини \mathbf{P} на кластери полягає у знаходженні такого розбиття $\mathbf{C} = \{C_1, C_2, \dots, C_k\}$, щоб образи у середині кожного кластера були максимально подібними, а образи з різних кластерів максимально відрізнялися один від одного [1, 2, 4]. Отже, оптимальне розбиття на кластери повинно мати такі властивості:

- кожен кластер повинен мати хоча б один образ

$$C_i \neq \emptyset \forall i \in \{1, 2, \dots, k\}; \quad (1)$$

- кожен образ повинен належати лише одному кластеру

$$C_i \cap C_j = \emptyset \forall i \neq j, i, j \in \{1, 2, \dots, k\}; \quad (2)$$

- усі образи повинні бути рознесені по кластерах

$$\bigcup_{i=1}^k C_i = \mathbf{P}. \quad (3)$$

Отже задача кластерного аналізу зводиться до оптимізаційної задачі, яка вимагає побудови і дослідження певних критеріальних функцій. Формально таку задачу можна подати як пошук оптимального розбиття \mathbf{C}^* з усієї множини можливих розбиттів на кластери:

$$\mathbf{U} = \{\mathbf{C}^1, \mathbf{C}^2, \dots, \mathbf{C}^{N(n,k)}\}, \quad (4)$$

де n – кількість образів; k – кількість кластерів; $N(n, k) = \frac{1}{k!} \sum_{i=1}^k (-1)^i \binom{k}{i} (k-i)^i$ – кількість можливих розбиттів на кластери за вибору різних варіантів кількості кластерів.

Аналіз вирішення проблеми

В n -вимірному евклідовому просторі відстань між двома точками визначається так:

$$D(\mathbf{a}, \mathbf{b}) = \|\mathbf{a}, \mathbf{b}\| = \sqrt{(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b})} = \sqrt{\sum_{i=1}^d (a_i - b_i)^2}, \quad (5)$$

де \mathbf{a}, \mathbf{b} – d -вимірні вектори.

Середньоквадратична відстань між точкою \mathbf{X}_i і множиною \mathbf{M}_j за умови, що $\mathbf{X}_i \notin \mathbf{M}_j$:

$$\overline{D^2(\mathbf{X}_i, \mathbf{M}_j)} = \frac{1}{n} \sum_{l=1}^n \sum_{k=1}^d (x_{i,k} - x_{l,k}^j), \quad (6)$$

де $x_{i,k}$ – k -та компонента вектора \mathbf{X}_i ; $x_{l,k}^j$ – k -та компонента вектора $\mathbf{X}_l \in \mathbf{M}_j$; n – кількість точок в множині \mathbf{M}_j .

Середньоквадратична відстань між усіма точками множини \mathbf{M} :

$$\overline{D^2(\mathbf{X}_i, \mathbf{X}_j)} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \sum_{k=1}^d (x_{i,k} - x_{j,k})^2, \quad (7)$$

де $\mathbf{X}_i, \mathbf{X}_j$ – d -вимірні вектори; n – кількість точок у множині \mathbf{M} .

Використання таких розрахункових формул безпосередньо під час оцінювання результатів кластеризації вимагає значних обчислювальних затрат. Тому актуальним є розроблення такого алгоритмічного забезпечення для оцінки результатів процесу кластеризації, яке уможливить отримати задовільні результати за незначних обчислювальних затрат.

Побудова алгоритму

Для побудови критеріальних функцій оцінки результатів процесу кластеризації використовуватимемо індекси оптимальності.

Застосування індексів оптимальності має на меті оцінити два параметри:

- подібність образів у межах усіх кластерів розбиття (мінімізація);
- відстань між кластерами розбиття (максимізація).

Загальну блок-схему алгоритму оптимального розбиття на кластери показано на рисунку. Робота алгоритму полягає в послідовному переборі методів кластеризації та апробації кількох варіантів вибору початкових параметрів роботи алгоритму для кожного індексу оптимальності. Оптимальним вважається таке розбиття на кластери, для якого отримане оптимальне значення хоча б одного індексу оптимальності.

Для оцінювання оптимальності результатів кластеризації використовується «силуетний» індекс, індекс Дана, індекс Деві-Болдвіна [5–7].

Під час розрахунку «силуетного» індексу для кожного кластера $C_j (j=1,2,\dots,k)$ і кожного образу кластера $X_i \in C_j (i=1,2,\dots,m)$ обчислюється величина $s(i)$ відома як «силуетний» індекс:

$$s(i) = \frac{B(i) - A(i)}{\max\{A(i), B(i)\}}, \quad (8)$$

де $A(i)$ – середня відстань між образом $X_i \in C_j$ і усіма образами кластера C_j ; $B(i)$ – мінімальна відстань між образом $X_i \in C_j$ і образами $X_l \in C_l (l=1,2,\dots,k), l \neq j$.

Величина $s(i) \in [-1;1]$. Якщо величина $s(i)$ близька до 1, тоді образ X_i оптимально віднесений до належного кластера (рисунок).

«Силуетний» індекс для кластера C_j обчислюється за такою формулою:

$$S_j = \frac{1}{m} \sum_{i=1}^m s(i), \quad (9)$$

де m – кількість образів у кластері C_j .

Загальний «силуетний» індекс для оптимального розбиття на кластери обчислюється за такою формулою:

$$GS = \frac{1}{k} \sum_{j=1}^k S_j, \quad (10)$$

де k – кількість отриманих кластерів.

Оптимальною величиною вважається максимальна величина загального «силуетного» індексу.

Індекс Дана обчислюється за такою формулою:

$$D(U) = \min_{1 \leq i \leq k} \left\{ \min_{1 \leq j \leq k} \left\{ \frac{d(C_i, C_j)}{\max_{1 \leq l \leq k} (\Delta(C_l))} \right\} \right\}, \quad (11)$$

де $d(C_i, C_j)$ визначає відстань між кластерами C_i та C_j ; $\Delta(C_l)$ – визначає відстань між усіма образами кластера C_l ; k – кількість отриманих кластерів.

Оптимальною величиною вважається максимальна величина індексу Дана.

Індекс Деві-Болдвіна обчислюється за такою формулою:



Загальна блок-схема алгоритму оптимального розбиття на кластери

$$DB(U) = \frac{1}{k} \sum_{j=1}^k \sum_{i=1}^k \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\}, \quad (12)$$

де – $\Delta(C_i)$, $\Delta(C_j)$, $d(C_i, C_j)$ визначають ті самі параметри, що і в «силуетному» індексі.

Оптимальною вважається мінімальна величина індексу Деві-Болдвіна.

Алгоритмічна оцінка оптимальності результатів кластеризації за критерієм відстані реалізована в автоматизованій системі кластерного аналізу [8; 9].

1. Ту Дж., Гонсалес Р. *Принципы распознавания образов.* – М.: Мир, 1978. 2. Дюран Б., Оделл П. *Кластерный анализ.* – М.: Статистика, 1977. 3. Мандель И.Д. *Кластерный анализ.* – М.: Финансы и статистика, 1988. 4. Барселян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. *Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP.* – СПб.: БХВ-Петербург, 2007. 5. Rousseeuw P.J. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics.* – 1987. – Vol.20. – P.53–65. 6. Bezdek J.C., Pal N.R. *Some new indexes of cluster validity. // IEEE Transactions on Systems, Man and Cybernetics.* – 1998. – Vol.28. – P.301–315. 7. Dunn J. *Well separated clusters and optimal fuzzy partitions // Journal of Cybernetics.* – 1974. – Vol.4. – P.95–104. 8. Stekh Y., Fajsal M.E. Sardieh, Lobur M. *System for a cluster analysis. – Proceedings of the Xth International Conference “Modern Problems of Radio Engineering, Telecommunications, and Computer Science” // IEEE TCSET’2010.* – Lviv–Slavske. – 2010. – P.236 9. Stekh Y., Fajsal M.E. Sardieh, Kernytskyy A., Nykyforchyn R. *Dialog graphical system of classification with the help of distance function. – Proceedings of the XVI Ukrainian-Polish Conference on “CAD in Machinery Design. Implementation and Education Problems” // CADMD’2008.* – Lviv, 2008. – P.88–89.

УДК 004.932

М.В. Олексів

Національний університет “Львівська політехніка”,
кафедра електронно-обчислювальних машин

ЦИФРОВЕ ОПРАЦЮВАННЯ ВІЗУАЛЬНИХ ЗОБРАЖЕНЬ З МЕТОЮ ЕФЕКТИВНОГО ВІДДІЛЕННЯ ОБ’ЄКТІВ ВІД ФОНУ

© Олексів М.В., 2010

Розглянуто спосіб цифрового опрацювання візуальних зображень з метою ефективного відділення об’єктів від фону. Спосіб полягає у специфічному поєднанні операцій фільтрації і вирівнювання гістограми зображення. Полегшено відділення об’єктів від фону на зображеннях, отриманих за різних природних умов.

Ключові слова – цифрове опрацювання зображень, відділення об’єктів від фону, фільтрація зображень.

The digital visual images processing method for effective objects extraction from the background is considered. The method consists in a specific combination of filtering operations and histogram equalization. It facilitates objects extraction from background in the images, which are obtained in different environmental conditions.

Keywords – digital image processing, objects extraction, image filtration.

Вступ

За цифрового опрацювання зображень сцени, отриманих за допомогою сенсорів в складних природних умовах, виникає проблема отримання таких зображень, що придатні для ефективного