УДК 519.688

К.т.н., ст. викладач Сирота С.В., студентка Береговська Х.В.

Національний технічний університет України
«Київський політехнічний інститут»

ПОРІВНЯЛЬНИЙ АНАЛІЗ РОБОТИ АЛГОРИТМУ *k*-СЕРЕДНІХ ТА АГЛОМЕРАТИВНОЇ ІЄРАРХІЧНОЇ КЛАСТЕРИЗАЦІЇ

Abstract

Sergiy V. Syrota, PhD, senior teacher; Cristina Beregovska, student

Comparative analysis of k-means algorithm and Aglomeration Hierarchical Clustering

This paper concerns the task of clustering's algorithm researching. It gives short overview of two clustering methods – k-means and Aglomeration Hierarchical Clustering, describes their advantages and disadvantages.The ways of methods improvement based on reserch results are introduced.

Вступ

Кластерний аналіз – це багатомірна статистична процедура, яка класифікує об'єкти або спостереження в однорідні групи. Набір усіх досліджуваних об'єктів розподіляється по підкласах, які називаються кластерами, класами, скупченнями або таксонами. Кластерний аналіз включає набір методів, які мають свої особливості та переваги. Областю застосування кластерного аналізу може бути як біологія (наприклад для розбиття тварин на різні види, з метою опису відмінностей між ними), так і економічне прогнозування. Він використовується при сегментції зображень, у фінансових задачах, а також в інших задачах, де потрібен поділ множини об'єктів на кластери, згідно характеристик цих об'єктів.

В даній статті розглядається неієрархічний метод *k*-середніх та порівнюється його робота з методом агломеративної ієрархічної кластеризації Ланса Уільямса. Досліджуються переваги та недоліки алгоритмів, робляться висновки щодо оптимізації їх застосування.

Постановка задачі

Задача полягає в дослідженні ієрархічних та неієрархічних методів кластеризації на прикладі алгоритмів *k*-середніх та агломеративної

кластеризації, виявлення особливостей цих методів, переваг та недоліків їх застосування на конкретному прикладі, а також знаходження шляхів удосконалення цих методів.

В якості прикладної області для даної задачі взято таблицю характеристик накопичувачів для збереження даних в інформаційних системах. До цих накопичувачів належать: твердотільні носії інформації SSD, RAID мисив SSD, магнітооптичні носії, DVD, Blu-ray, UDO (Ultra Density Optical), HDD, RAID масив HDD, MT.

Термінологія

Кластер (англ. *cluster* - скупчення) – об'єднання декількох однорідних елементів, яке може розглядатись як самостійна одиниця, котра володіє певними властивостями.

Кластеризація – задача статистичного аналізу, котра передбачає розбиття заданої вибірки об'єктів (ситуацій) на підмножини, котрі не перетинаються (кластери), так, щоб кожен кластер складався із схожих між собою об'єктів, а об'єкти різних кластерів суттєво відрізнялись.

Опис алгоритмів

Алгоритми кластеризації можна розділити на ієрархічні та неієрархічні. Основна особливіть ієрархічних методів полягає в тому, що вони будують певну ієрархію груп, розбиваючи множину досліджуваних об'єктів. Неієрархічні алгоритми при поділі на кластери керуються певною цільовою функцією, значення якої намагаються мінімізувати [1].

1. Постановка задачі кластеризації для всіх алгоритмів полягає в наступному:

   Дано: X- простір об'єктів; $X^l=\{x_i\}_{i=1}^l$ – вибірка елементів;

   $d: X \times X \to [0, \infty)$ – функція відстані між об'єктами.

   Знайти: Y – множину кластерів і $a: X \to Y$ – алгоритм кластеризації такий, що кожен кластер складається з близьких між собою об'єктів, об'єкти різних кластерів суттєво відрізняються [1].

2. Агломеративна ієрархічна кластеризація. Алгоритм Ланса – Уільямса:

   2.1. Розгянемо всі кластери, як одноелементні масиви:

   $t := 1$, $C_t = \{\{x_1\},...,\{x_l\}\}$, $R(\{x_i\},\{x_j\}) := d(x_i, x_j)$.

   2.2. Для всіх t=2,…,l (t – номер ітерації) знаходимо в C_{t-1} два найближчі кластери: $(U,V) := \arg\min_{U \neq V} R(U,V)$, $R_t := R(U,V)$, $W = U \cup V$.

2.3. З'єднуємо в один кластер: $C_t := C_{t-1} \cup \{W\} \setminus \{U,V\}$.

2.4. Для всіх $S \epsilon C_t$ обчислюємо $R(W,S)$ по формулі Ланса-Уільямса:

$$R(U \cup V,S) = \alpha_U \cdot R(U,S) + \alpha_V \cdot R(V,S) + \beta \cdot R(U,V) + \gamma \cdot |R(U,S) - R(V,S)|,$$

де α_U, α_V, β, γ - числові параметри, значення яких залежать від метод вибору *R(W, S)* (по ближньому чи дальньому сусіду, груповій середній відстані, відстані між центрами чи відстані Уорда) [2].

Її застосування полягає у тому, щоб визначити відстань *R(W, S)* між кластерами $W = U \cup V$ та *S*, знаючи відстані *R(U, S),R(V, S), R(U, V).*

3. Алгоритм *k*-середніх.

- задаємо *k* об'єктів, котрі будуть еталонами. Ці точки вважаються центрами ваги кластерів на першому кроці процедури. Кожному еталону присвоюється порядковий номер *L* (*L*=1,2,3…,*k*).

- з решти (*n*-*k*) об'єктів вибирається точка x_i (*i*=1,2,..*n*) з координатами (x_{i1}, x_{i2},..,x_{im}) і перевіряється до якого з еталонів вона знаходиться найближче. Для цього використовується евклідова метрика. Об'єкт, що перевіряється, приєднується до того центру ваги (еталону), якому відповідає мінімальна відстань. Еталон заміняється новим, перерахованим з урахуванням нової точки, а його статична вага (кількість об'єктів, що входять у даний кластер) збільшується на 1. Аналогічно вибираємо і перевіряємо для решти з (*n*-*k*) точок.

Щоб отримати стійкий розподіл об'єктів по кластеру всі точки x_1, x_2,..,x_n знову приєднуються до отриманих в результаті попереднього кроку центрів ваги. Новий розподіл об'єктів порівнюється з попереднім. Якщо вони співпадають, то робота алгоритму припиняється. В іншому випадку цикл повторюється [3].

Кінцевий розподіл, зазвичай, має центри ваги, що не співпадають з початковими еталонами. Їх можна позначити через $c_1, c_2, \ldots, c_k$. При цьому кожна точка x_i буде відноситись до такого кластеру *L*, для котрого виконується вимога: $d(x_j,c_L) = \min\limits_{1 \le L \le k} d(x_j,c_L)$.

4. Результати дослідження

Проведене дослідження цих методів виконувалось на основі характеристик накопичувачів для збереження даних. Вхідні дані – таблиця характеристик накопичувачів (табл. 1) .

Таблиця 1

Вхідні дані для дослідження

	Ємність	Швидкість доступу	Строк збереження інформації	Ціна (за 100 шт)	Надійність
Твердотільні носії інформації SSD	400 Гб	270 Мб/с	7 років	300 у.о	2 років
RAID масив SSD	…	…	…	…	…
МТ	стандарт LTO-5 1,4 Тб	обміну нестиснутими/стиснутими даними – 140 / 280 Мб/с	21 рік	250у.о	3,5 роки

Результатом роботи є поділ на кластери. Взявши за характеристику подібності термін збереження інформації, отримано такі розбиття: згідно методу *k*-середніх до першої групи належать: Blue-Ray(1), UDO(2) та магнітооптичні носії(3), до другої: RAID массив HDD(4), MT(4), RAID массив SSD(6), до третьої: DVD(7) , трердотільні носії інформації SSD(8), HDD(9). Використовуючи агломеративну ієрархічну кластеризацію, отримано схожі результати за винятком того, що накопичувачі МТ формують окремий одноелементний кластер (рис.1).

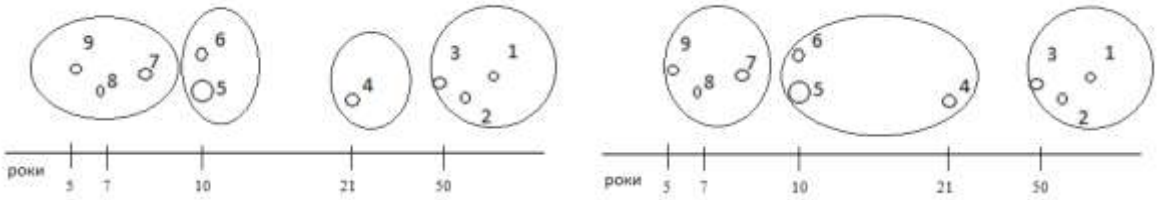

Рис.1. Графічне зображення результатів кластеризації

Певні відмінності у результатах є наслідком того, що для даної задачі немає чіткої оцінки якості рішення, а умова припинення поділу на кластери часто є неоптимальною.

В результаті порівняльного аналізу асимптотичної складності на основі проведеного дослідження зроблено висновки про швидкість виконання кластеризації кожним із методів. Згідно цього аналізу швидкість роботи алгоритму Ланса-Уільямса предстравляє собою функцію, котра кубічно залежить від кількості досліджуваних об’єктів, а ця залежність для методу *k*-середніх є такою, котра наближується до лінійної. З цього робимо висновок, що для невеликої вибірки (приблизно 20 об’єктів згідно досліду) добре підходять обидва алгоритми, але якщо масив даних більший (наприклад, 100-200 об’єктів і більше) набагато оптимальніше використовувати алоритм *k*-середніх. Проте варто

зазначити, що реалізація методу *k*-середніх є складнішою, причиною чого є підбирання коректних початкових наближень для центрів ваги.

Висновки

Згідно результатів проведеного дослідження можна зробити висновок, що як метод *k*-середніх, так і метод агломеративної ієрархічної кластеризації можуть вдало використовуватись для проведення кластеризації в різних прикладних областях, причому результати цієї кластеризації будуть близькими.

Проблема ефективності методу Ланса-Уільямса полягає у трудомісткості пошуку ближніх кластерів на другому кроці:

$(U,V) \coloneqq \arg\min_{U \neq V} R(U,V)$, що збільшує час пошуку результату, а значить, зменшує ефективність роботи алгоритму.

Для підвищення ефективності можна запропонувати перебирати лише близькі пари:

$(U,V) \coloneqq \arg\min_{R(U,V) \leq \delta} R(U,V)$ та періодично збільшувати параметр δ.

Основним недоліком методу *k*-середніх є те, що потрібно заздалегідь задавати *k* - кількість кластерів та еталонів, що не завжди можливо зробити раціонально. Метод є дуже чутливим до цих початкових наближень значень центрів. Для усунення цієї проблеми можна застосовувати методику поступового збільшення значення числа кластерів.

Література

1. *Мандель И.Д.* Кластерный анализ / И.Д. Мандель. – М. : Финансы и статистика, 1988. – 176 с
2. *Гитис Л.Х.* Кластерный анализ в задачах классификации, оптимизации и прогнозирования / Л.Х. Гитис. – М. : МГГУ, 2001. – 104 с.
3. *Березін Б. О., Качанов П. Т., Циганок В. В., Андрійчук О. В.* Підтримка прийняття рішень при побудові систем довготермінового зберігання інформації.// Проблеми розвитку інформаційного суспільства: матеріали Міжнародного форуму. -2009.- С.145-153.
4. *Гитис Л.Х.* Статистическая классификация и кластерный анализ / Л.Х. Гитис. – М. : МГГУ, 2003. – 157 с.