Дємєнтьєв А.В.
студент,
Національний технічний університет України
«Київський політехнічний інститут»

МОДИФІКАЦІЯ ІЄРАРХІЧНОГО АЛГОРИТМУ КЛАСТЕРИЗАЦІЇ З ДВОНАПРАВЛЕНИМ ПОШУКОМ

Кастеризація є одним із методів аналізу та обробки даних, який широко застосовується в археології, медицині, психології, хімії, біології, філології, маркетингу, соціології та інших дисциплінах. Головна задача при проведенні кластеризації полягає у розбитті даних на окремі підгрупи – кластери, що схожі за певними ознаками. Ця задача є актуальною через необхідність систематизації інформації, об'єми якої постійно зростають підчас інформатизації різних галузей діяльності людства.

Зростання попиту на застосування методів кластерного аналізу породжує зростання їх кількості. Відомі методи кластеризації можливо відносити чи до ієрархічнх чи до неієрархічних. У неієрархічних методах, як правило, задають кількість кластерів як параметр алгоритму (k-means, PAM кластеризація та ін.) або використовують деякі алгоритмічні процедури знаходження їх кількості (CLOPE, карти Кохонера та ін.). Ієрархічна кластеризація виконується чи за допомогою послідовного об'єднання кластерів (агломеративні процедури) чи навпаки за допомогою послідовного розбиття кластерів (дивізийні процедури). Однак, не зважаючи на велику кількість можливих алгоритмів, результати більшості з них залежать від суб'єктивного фактору (наприклад, від початкових налаштувань заданих дослідником) та не дають оптимального варіанту розбиття.

Одним з алгоритмів, що використовують ефективний підхід для пошуку найкращої кластеризації та пошуку інформативних ознак є алгоритм Саричевої Л.В [1]. Суть даного алгоритму полягає у поєднанні агломеративного та дивізимного алгоритму кластеризації та знаходженні найбільш подібного розбиття при рівній кількості кластерів. Це дає змогу побудувати ядра кластеризації і маніпулювати даними що не ввійшли до кластерів. Характерною особливістю ієрархічної двонаправленної кластеризації є автоматичне визначення числа кластерів при досягненні максимальної схожості між кластеризаціями отриманими за допомогою дивізимного та агломеративного алгоритмами. Завдяки цьому вирішується проблема пошуку кількості кластерів та знаходиться найкраща кластеризація.

Однак, двонаправлений ієрархічний алгоритм залишає поле для інтерпретації, так як вибір дивізимного та агломеративного алгоритму та вибір критеріїв порівняння відстані між точками залишаються суб'єктивними. Також алгоритм є доволі громіздким для обчислення великої кількості даних.

В нашому дослідженні, була поставлена задача порівняння різних функціональних ієрархічних алгоритмів, автоматизації попередньої обробки даних з метою пошуку кращої функціональної кластеризації та зменшення складності обчислень шляхом встановлення границь пошуку.

Алгоритм складається з наступних етапів:

1. Передобробка даних (нормування та центрування даних).

2. Проведення двонаправленої кластеризації. Двосторонній алгоритм розраховує розбиття для агломеративної кластеризації, після чого проводиться розбиття дивізимним алгоритмом, доки кількість кластерів не буде дорівнювати кількості кластерів при агломеративному.

3. Розрахунок відстані між побудованими розбиттями кластерів. Розбиття вважаються тим кращими, чим більше відстані наближаються до одиниці. При відстані рівній одиниці обидва розбиття співпадають. Стан з більшою відстанню запам'ятовується і вважається найкращим.

Викладений алгоритм є циклічним і повторюється до тих пір поки не дійде до граничного значення заданого користувачем. За отриманим результатом розраховуються ядра кластерів. В них потрапляють елементи, що входили до спільних кластерів. Спільний кластер – це кластер, в якому кількість однакових елементів в обох кластеризаціях є найбільшою. Наступним кроком всі елементи, що не потрапили до ядер, додаються до кластерів за принципом найменшої відстані між даною точкою та точкою в кластері.

В пропонованій роботі були реалізовані дивізимний алгоритм пошуку найменшої відстані між точками в кластері та поза ним та агломеративний алгоритм KRAB. Мірою відстані між точками було вибрано Евклідову метрику.

Алгоритм був реалізований на мові програмування C# з використанням бібліотек пакету.NET. Було проведено дослідження на вибірці даних $X = X\{x_1, x_2, x_3, x_4\}$, яка має назву «Іриси Фішера». За попередніми даними у вибірці виділяють три кластери. При побудові за допомогою двостороннього алгоритму були утворені три кластери, що задовольняє вхідним даним.

Висновки. Запропоновано модифікацію двонаправленого алгоритму кластеризації, який відрізняється підвищеною швидкістю розбиття на групи. Якість кастеризації було задовільно протестовано на стандартному наборі даних «Іриси Фішера».

Список використаних джерел:

1. Саричева Л. В. Єдиний підхід до класифікації методів кластеризації та методів вибору інформативних ознак // Актуальні проблеми автоматизації та інформаційних технологій. Том 6. Збірник наук. праць. Дніпропетровськ: Вид-во «Навчальна книга», 2002. – С. 137-144.

2. Воронцов К. В. Лекції по алгоритмам кластерізації і багатовимірного шкалювання. – www.ccas.ru/voron/download/Clustering.pdf

3. Огляд алгоритмів кластеризації даних [Електронний ресурс]. – 2010. – Режим доступу до ресурсу: https://habrahabr.ru/post/101338/.

4. Жамбю М. Ієрархічний кластер-аналіз та відповідності. – М.: Финансы и статистика, 1988. – 345 с.