УДК 004.42 519.1 582.099 581.4

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ РІЗНИХ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ ДЛЯ ОБРОБКИ ДАНИХ ПОЛЬОВОГО ЕКСПЕРИМЕНТУ З ДОСЛІДЖЕННЯ РОДИННИХ ЗВ’ЯЗКІВ ЛІНІЙ СОНЯШНИКУ

Сіренко Р.В., студент, skriler222000@gmail.com, НУ «ЗП»
Терещенко Е.В., доцент, elina_vt@ukr.net, НУ «ЗП»

У контексті вивчення родинних зв’язків ліній соняшнику виникає проблема необхідності ефективної математичної обробки обширних даних, отриманих у ході польового експерименту. Зібрані дані мають великий обсяг та включають різноманітні параметри, що потребують точної обробки та аналізу. Основні проблеми створеного програмного застосунку виникають через обмежений функціонал існуючого алгоритму, який ґрунтується на повному переборі.

Реалізований алгоритм обмежує ефективність та точність аналізу, що призводить до проблем у функціонуванні програмного застосунку. Основні недоліки включають обмежену ефективність, недостатню точність порівнянь і відсутність альтернативних алгоритмів. Мета задачі полягає у вирішенні цих проблем через розробку та впровадження альтернативних алгоритмів, використання методів кластеризації та підвищення точності порівнянь.

Постановка задачі полягає у необхідності проаналізувати різні алгоритми кластеризації та визначити найбільш ефективний для обробки даних польового експерименту щодо родинних зв’язків ліній соняшнику. Процес включає вивчення теоретичних основ алгоритмів та експериментальну оцінку їх ефективності. Результатом буде обрання найбільш підходящого алгоритму для подальшого вдосконалення програмного застосунку.

Методи формування кластерів у кластерному аналізі можуть базуватися на відстані між об'єктами, щільності ділянок у просторі даних, інтервалах або конкретних статистичних розподілах. Вибір конкретного методу залежить від характеристик даних та мети використання результатів. Кластерний аналіз - це ітераційний процес, оскільки доводиться експериментувати з методами обробки даних та параметрами моделі, щоб досягти бажаних властивостей результатів [1].

Рішення задачі кластеризації може бути неоднозначним і ускладненим з кількох причин. По-перше, відсутність універсального критерію, який однозначно визначав би якість кластеризації, призводить до того, що різні критерії оцінки можуть привести до різних результатів, ускладнюючи вибір оптимального рішення. По-друге, встановлення кількості кластерів здійснюється на основі суб'єктивних оцінок, що додає ступінь невизначеності до процесу кластеризації. По-третє, вибір метрики визначається експертом, і це може призводити до різних інтерпретацій та варіацій у кінцевих результатів.

Алгоритми кластеризації можна класифікувати за різними критеріями, і одним з основних є тип кластеризації. Тип кластеризації може бути жорстким або м'яким [2]. У жорсткій кластеризації кожному об'єкту присвоюється тільки один

кластер, а об'єкт може належати або до одного кластера, або жодного. При м'якій кластеризації кожному об'єкту може бути присвоєна ймовірність належності до кожного кластера, дозволяючи об'єкту належати до декількох кластерів одночасно з різною ймовірністю.

Інший важливий критерій при класифікації алгоритмів кластеризації - це метод визначення кластерів, який може бути ієрархічним, неієрархічним або гібридним. Ієрархічні алгоритми будують ієрархію кластерів, поділяючись на агломеративні (починають з одного кластера та поступово їх об'єднують) та дивізійні (починають з набору кластерів та поступово їх ділять). Неієрархічні алгоритми визначають кластери без побудови ієрархії, поділені на центровані, за щільністю та за формою. Гібридні алгоритми використовують комбінацію ієрархічних та неієрархічних методів кластеризації.

Таблиця 1 – Порівняння алгоритмів

Алгоритм	Обчислювальна складність (у більшості випадків)	Гнучкість щодо форми кластерів	Вибір кількості кластерів	Простота реалізації	Чутливість до шуму
K-середніх	$O(n^2 * k)$	Низька	Користувачем	Простий	Нестійкий
НАС	$O(n^3 * \log(n))$	Висока	Автоматично	Середній	Стійкий
DBSCAN	$O(n * \log(n))$	Висока	Автоматично	Середній	Стійкий
Метод спектральної кластеризації	$O(n^2 * k)$	Висока	Користувачем	Складний	Стійкий
OPTICS	$O(n * \log(n))$	Висока	Автоматично	Складний	Стійкий

Висновки: На основі проведеного аналізу різних алгоритмів кластеризації з метою вибору найбільш ефективного для обробки даних польового експерименту щодо родинних зв’язків ліній соняшнику, було встановлено, що алгоритм DBSCAN є оптимальним вибором.

DBSCAN вирішує проблеми, які виникають при використанні повного перебору, забезпечуючи високу обчислювальну ефективність, точність порівнянь та шумостійкість. Його обчислювальна складність робить його практичним для обробки невеликих об'ємів даних, а простота реалізації дозволяє легко інтегрувати його у програмний застосунок. Особливо важливою перевагою DBSCAN є його здатність автоматично визначати кількість кластерів без необхідності вручну задавати цей параметр.

Список використаних джерел

1. Бойко В.В. Алгоритми кластеризації [Текст] / Бойко В.В, Гаврилов Ю.Ф., Завадський І.К. – К.: Наукова думка, 2002. – 256 с.
2. Aggarwal, C. C. Data clustering: algorithms and applications [Текст] / Aggarwal, C. C., Reddy, C. K. – CRC Pres, 2013. – 311 с.