

*Карижська Поліна Григорівна
студентка I курсу магістратури
Спеціальності «ІІІ Математика»*

*Наукові керівники:
доктор фізико-математичних наук, професор
Кондратьєв Юрій Григорович,
старший викладач
Горбачук Василь Олександрович*

«k-середніх: переваги та недоліки методу кластеризації даних»

На сьогодні глобальна комп'ютеризація та швидкий розвиток інформаційно-комунікаційних технологій (ІКТ) обумовлюють значне щорічне збільшення об'ємів інформації. Часто можна почути думку, що однією із головних сучасних проблем у сфері ІКТ є збереження такого об'єму інформації. Проте, на нашу думку, найбільша проблема полягає не в збереженні даних, а в їх обробці. Таким чином, потрібно мати дієві механізми для спрощення даних та їх подальшого аналізу. Одним із таких методів є кластерний аналіз.

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. Побудова однієї загальної моделі для всіх даних, що опрацьовуються, може бути досить складною задачею. Тому аналітику часто набагато легше виділити групи схожих об'єктів, вивчити їх особливості і побудувати для кожної групи окрему модель. Так, наприклад, у маркетингових дослідженнях часто виділяють групи клієнтів, покупців, товарів і розробляють для кожної з них окрему стратегію. Також областю застосування кластерного аналізу можуть бути: біологія, філологія, соціальні науки, економічне прогнозування тощо. Метод k-середніх використовується до всього іншого, при сегментації зображень, обробці живої мови, а також в інших задачах, де потрібен поділ множини об'єктів на кластери, згідно характеристик цих об'єктів.

k-середніх – це найбільш розповсюджений і найліпше вивчений з-поміж усіх інших методів кластеризації даних. Дослідженням цього методу, його недоліками, перевагами та вдосконаленням займалися Штейнхаус Х., Лойд С., Макквін Д., Горбань А.Н., Зінов'єв А.Й., Артур Д., Васильвіцький С., Ткаченко О.[1].

Кластеризація методом k-середніх – впорядкування множини об'єктів в порівняно однорідні групи.

Припустимо, дослідник має гіпотезу щодо кількості кластерів даних – N . У такому випадку, можна створити рівно N кластерів так, щоб вони були настільки різні, наскільки це можливо. Призначення методу k-середніх якраз і полягає в розв'язанні завдань цього типу. Гіпотеза може базуватись на теоретичних міркуваннях, результатах попередніх досліджень або здогадках [3]. Також слід зазначити, що розбиття на різну кількість кластерів (тобто розв'язання задачі декілька разів для різних N) дає змогу порівняти якість і коректність отриманих рішень.

Цього можна досягти, почавши дослідження спочатку з N випадково обраних кластерів, а потім змінювати приналежність об'єктів до них, щоб мінімізувати мінливість всередині кластерів і максимізувати мінливість між кластерами. Алгоритм випадковим чином в просторі призначає центри майбутніх кластерів. Потім обчислює відстань між центрами кластерів і кожним об'єктом, і об'єкт приписується до того кластеру, до якого він найближчий. Після розподілу всіх об'єктів, алгоритм обчислює середні значення для кожного кластера. Цих середніх буде стільки, скільки використовується змінних для проведення аналізу – k . Набір середніх являє собою координати нового положення центру кластера. Такий процес повторюється доти, поки центри тяжіння не перестануть «мігрувати» в просторі [3].

Поширеність методу k-середніх зумовлено його головними перевагами:

- простотою у використанні;
- гнучкістю;
- швидкою збіжністю;

- зрозумілістю;
 - можливістю перевірки статистичної значимості відмінностей між виділеними кластерами.
- Але метод суттєво обмежується такими недоліками:

- результати кластеризації за методом k-середніх значною мірою залежать від вибору початкової конфігурації центроїдів (для усунення цієї проблеми можна застосовувати методику поступового збільшення значення числа кластерів) [1];
- кількість кластерів повинна бути заздалегідь визначена дослідником;
- робота алгоритму суттєво уповільнюється під час кластеризації великих обсягів даних (можливим вирішенням даної проблеми є використання вибірки даних);
- алгоритм може сходитися до локального мінімуму цільової функції [1];
- алгоритм чутливий до викидів, які можуть викривлювати середнє;
- порушення умови зв'язності елементів одного кластера [2].

Оскільки цей метод має значні недоліки, почали розвиватись його модифікації та нечіткі аналоги (англ. fuzzy k-means methods). Вони відрізняються тим, що на першій стадії алгоритму допускається належність елемента до декількох кластерів з різним рівнем приналежності. На відміну від k-середніх, k-медіан для визначення центроїдів використовує медіану. Ця зміна робить алгоритм стійкішим до аномальних значень даних. У свою чергу, алгоритм g-середніх будує кластери з розподілом даних, що прямує до нормального – це усуває проблему невизначеності при початковому виборі кластерів.

Отже, підсумовуючи сказане, k-середніх є найбільш дослідженим і часто застосовуваним методом кластеризації даних. Проте, маючи перелік беззаперечних переваг, k-середніх має також велику кількість недоліків, які не дають йому можливості претендувати на звання простого і одночасно універсального методу кластеризації великих об'ємів даних. Саме тому в подальших дослідженнях ми більш докладно вивчимо його модифікації, а також інші методи кластеризації даних, які покликані допомогти переступити обмеження методу k-середніх.

Список використаної літератури:

1. Ткаченко О. М., Біліченко Н. О., Грійо Тукало О. Ф., Дзісь О. В. Метод кластеризації даних на основі послідовного запуску k-середніх з обчисленням відстаней до активних центроїдів. *Реєстрація, збирання та обробка даних*, 2012. Т. 14, № 1. С.25–34.
2. Сирота С. В., Береговська Х. В. Порівняльний аналіз роботи алгоритму k-середніх та агломеративної ієрархічної кластеризації. *Прикладна математика та комп'ютинг*, 2011. Вип. 70. С.311–315.
3. Теоретичні основи розроблення графіків навантажень. URL: <http://mpe.kmu.gov.ua/minugol/doccatalog/document?id=245201761> (Дата звернення: 23.03.2021).