Технічні науки

Юр Артем Сергійович
студент,
Національний технічний університет України «КПІ»
м. Київ, Україна
Наукові керівники:
Сердаковський В. С.
старший викладач кафедри біомедичної кібернетики,
Національний технічний університет України «КПІ»
Носовець О. К.
старший викладач кафедри біомедичної кібернетики,
Національний технічний університет України «КПІ»

РОЗРОБКА ПОКРАЩЕНОГО ЩІЛЬНІСНОГО АЛГОРИТМУ КЛАСТЕРИЗАЦІЇ

Аналіз даних є однією із надзвичайно важливих напрямків людської діяльності на сьогодні, особливо використання найсучасніших знань в області штучного інтелекту. Звідси і почала свій розвиток дисципліна інтелектуального аналізу даних.

Серед ряду задач, які розглядаються дисципліною Data Mining, є кластеризація даних.

Існує багато алгоритмів, які виконують задачу кластеризації вхідних даних, але в такому випадку виникає питання вибору між цими алгоритмами. Проблема складається в тому, що дані можуть володіти різними властивостями, а це в свою чергу призводить диференціації і придатності різних алгоритмів кластеризації.

Найбільшою проблемою існуючої реалізації щільнісного алгоритму кластеризації є те, що алгоритм погано працює з даними де щільність кожного кластеру має різне значення, бо як було описано в першій частині,

при збільшенні параметру ε росте ймовірність утворення одного великого кластеру замість утворення одного додаткового [1, 2].

Також значних зусиль та часу при роботі із алгоритмом вимагає вибір значення для параметру ε і не завжди емпірично вдається коректно його підібрати [2].

Саме ці дві задачі було на меті вирішити при розробці удосконаленого щільнісного алгоритму кластеризації даних.

При дослідженні роботи алгоритму було встановлено, що однозначної відповіді для задання початкового значення ε із існуючими методами – немає змоги. Це в першу чергу пов'язано із тим, що кластери містять різну щільність, а тому цей параметр може приймати різне значення в залежності від досліджуваної ділянки даних [1]

Оскільки саме параметр ε в більшій мірі впливає на роботу алгоритму і встановлено, що він може приймати різні значення в залежності від досліджуваної області вхідних даних, було зроблено висновок, що на вхід алгоритму необхідно ітеративно передавати різні значення цього параметру, а отримані результати аналізувати.

Очевидно, що для певних інтервалів значень ε ми отримували різні результати, тому необхідно було продумати критерій, який би оцінював якість проведеної кластеризації.

Найбільш доцільним параметром для оцінки результату кластеризації стала кількість самих кластерів, які обов'язково задовольняють первинним вимогам алгоритму, а саме кількість об'єктів в кластері повинна бути більшою, аніж параметр *MinPts*, який за замовчуванням приймає значення на одиницю більше від кількості признаків даних [2].

Були проведені дослідження існуючих критеріїв та зупинились на оцінці, яка використовується в факторному аналізі, в методі головних компонент (PCA). Виконати дану оцінку можна за допомогою аналізу

власних чисел кореляційної матриці, яка була одержана від транспонованої матриці вхідних даних.

Після розрахунку власних чисел матриці, вони проходять через оцінку критерієм Кайзера, в якому $\lambda_i > 1;\ i = 1, \dots, n$, де λ_i – власне число вхідної транспонованої матриці даних. Кількість таких власних чисел і буде оціненою кількістю кластерів [3].

Варто також зауважити, що даний метод має свої обмеження пов'язані із властивостями використовуваних математичних методів. До них відносяться неможливість розрахунку рангу матриці для простого вектору (один рядок). Другим обмеженням є те, що кількість власних чисел не може перевищувати кількості признаків вхідних даних, в силу того, що ранг матриці також не перевищує даної кількості [3].

Тому дану оцінку можна використовувати, якщо вона приймає значення менше, аніж кількість признаків даних, в іншому випадку ми не можемо бути впевненими в вірності цього значення. Варто також зазначити, що використовуваних прикладах роботи алгоритму цього було достатньо та саме це питання повинне стати для покращення алгоритму в майбутньому.

В такому випадку після кожної ітерації виконання алгоритму проводилось порівняння із найкращим результатом та оціненою кількістю кластерів, якщо вона менша, аніж кількість признаків вхідних даних. Найкращий результат вважався той, який максимізував кількість кластерів при цьому мінімізуючи кількість точок, які вважались шумом. Тому даний крок звівся до класичної задачі max-min.

На допомогу прийшов метод розрахунку відстаней між найближчими сусідами *kNNdist*, окрім вхідних даних ми повинні також передати на вхід алгоритму кількість найближчих сусідів, які використовуються (в нашому випадку це мінімальна кількість об'єктів

необхідна для утворення кластерів). Таким чином, ми отримуємо матрицю дистанцій, яку згодом перетворюємо у вектор d [2].

Дослідивши властивості алгоритму ми з'ясували, що починаючи із найбільшої дистанції, в якості параметра ε, кількість утворюваних кластерів починає зростати, потім досягає свого піку, після чого починає знову скорочуватись. Тому було вирішено в процесі ітеративного переходу зменшувати значення ε. Причому доцільним виявилось використовувати 0,1% від початкового значення.

Емпіричним шляхом було встановлено, що для ініціалізації параметру ε доцільно використовувати:

$$\varepsilon = \mathrm{M}(d) + 2 * \sigma(d)$$

Також можна використовувати три стандартних відхилення замість двох, але у всіх досліджуваних випадках двох було цілком достатньо.

Важливо розуміти, що при некоректному підборі ε, тобто, якщо воно досить мале, то алгоритм може утворити значну частину шуму, навіть у вибірці, в якій спостерігається одноманітна щільність. З іншої сторони занадто велике значення ε може призвевсти до тривалих розрахунків, особливо на великих об'ємах даних.

Список використаних джерел:

1. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu Електроний ресурс: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

2. DBSCAN Revisited: Mis-Claim, Un-Fixability, and Approximation Junhao Gan, Yufei Tao. Електроний ресурс: http://www.cse.cuhk.edu.hk/~taoyf/paper/sigmod15-dbscan.pdf

3. Ким Дж.-О., Мьюллер Ч. У. «Факторный анализ: статистические методы и практические вопросы» / сборник работ «Факторный, дискриминантный и кластерный анализ»: пер. с англ.; Под. ред. И. С. Енюкова. — М.: «Финансы и статистика», 1989. — 215 с.