УДК 501

АДАПТИВНА КЛАСТЕРИЗАЦІЯ В ГАЛУЗІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Д. С. Хомініч[1, а], С. А. Смирнов[1, б]

[1] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут*

Анотація

Стаття присвячена дослідженню методів і алгоритмів кластеризації, яка широко використовується в системах інтелектуального аналізу даних. Особливий інтерес до методів аналізу даних виник у зв'язку з розвитком засобів збору і зберігання даних, що дозволив накопичувати великі обсяги інформації. Перед фахівцями постало питання про обробку даних, та перетворення їх в знання. Задача кластеризації дозволяє розділити на групи великі об'єми даних, а так само зменшити об'єми оброблюваної інформації. Можна скоротити розмір початкової вибірки, взявши один або кілька найбільш типових представників кожного кластера. Завдання кластеризації дуже добре підходить для виявлення шуму в даних, а саме виділення об'єктів, які не підходять за критеріями в жоден кластер. Виявлені об'єкти в подальшому обробляють окремо.

Вступ

У нашому сучасному інформаційному світі існують дуже великі об'єми інформації. Через це, дуже важко взяти щось корисне, бо усього життя не вистачить, щоб знайти дійсні і потенційно корисні дані в сучасних базах даних. Цим займається галузь знань, яка відноситься до обробки даних, що вивчає пошук і опис прихованих, нетривіальних і практично корисних закономірностей в досліджуваних даних яка називається інтелектуальний аналіз даних. Важлива відмінність інтелектуального аналізу даних від традиційних методик пов'язано з виявленням прихованих закономірностей в даних, а традиційні займаються оцінкою вже відомих закономірностей. Задачі інтелектуального аналізу даних [1]: класифікація – віднесення об'єктів до одного з заздалегідь відомих класів, кластеризація – це завдання розбиття множини об'єктів на групи, які називаються кластерами, в яких знаходяться схожі елементи, регресія - дослідження впливу одного або декількох незалежних змінних на залежну та задача пошуку асоціативних правил – пошук причинно-наслідкового зв'язку між подіями. На практиці, часто виникає задача первинного аналізу, коли про залежності в даних нічого невідомо. В таких умовах першою задачею аналізу, яку треба буде розв'язати, є виявлення внутрішньої структури в даних, на підставі якої можна буде формулювати більш детальні завдання про пошук залежності, що впливають на групування даних у вихідній множині. Це як раз і є задача кластеризації. Ми розглянемо основні методи кластеризації даних, а також їх модифікації.

Алгоритми кластеризації діляться на ієрархічні і неієрархічні. Ієрархічні алгоритми пов'язані з побудовою дендрограмм і діляться на агломеративні і дівізімні. Агломеративні – число кластерів зменшується, а у дівізімних збільшується.

Дендограма – це схема яка представляє собою дерево. Вона показує ступінь близькості окремих об'єктів і кластерів, а також наочно демонструє в графічному вигляді послідовність їх об'єднання або поділу. Щоб описати ступінь схожості, простір в якому знаходяться об'єкти, може бути прийнята скалярна метрика $\mathrm{d}(x, y)$ – це відстань між всякими двома об'єктами. Ця метрика повинна бути симетричною, невід'ємною, та відповідати рівнянню трикутника.

Основні методи кластеризації даних:

1) К-середніх.
2) Метод нечіткої кластеризації С-середніх.
3) Нейронна мережа Кохонена
4) ЕМ-алгорітм
5) Генетичний алгоритм

Чіткий поділ на кластери можливо тільки в ідеальних умовах, тому для вирішення реальних завдань частіше застосовуються нечіткі методи, в яких розбиття об'єктів виконується на частково перетинаються підмножини. В якості критерію схожості об'єктів вводиться міра близькості. Для кожного методу кластеризації вона завдається окремо, судячи по природі та типу наших даних.

Методи кластерного аналізу

Метод k-середніх – це метод кластерного аналізу, мета якого є поділ m спостережень на k кластерів, при цьому кожне спостереження відноситься до того кластеру, до центру (центроїду) якого воно найближче [2]. В якості міри близькості використовується

[а] dimaxomini4@gmail.com
[б] admin-sergsmir-ipt@lll.kpi.ua

евклідова відстань:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}.$$

Дія алгоритму таке,треба мінімізувати сумарне квадратичне відхилення точок кластерів від центрів цих кластерів:

$$V = \sum_{i=1}^{k}\sum_{x\in S_i}(x-\mu_i)^2, i = \overline{1,k}$$

де k – кількість кластерів,
S_i – отримані кластери,
μ_i – центри.

Метод нечіткої кластеризації С-середніх дозволяє розбити наявну безліч елементів потужністю N на задане число нечітких множин k. Метод нечіткої кластеризації С-середніх можна розглядати як вдосконалений метод k-середніх, при якому для кожного елемента розраховується ступінь його приналежності кожному з кластерів. В основі цього методу лежить теорія нечітких множин. Вхідні дані: масив об'єктів M, число кластерів c, параметр зупинки $\varepsilon > 0$.

1) Уточнення центрів кластерів за ступенями належності:
$$V_i = \frac{\sum_{k=1}^{M}\mu_{ki}^m * X_k}{\sum_{k=1}^{M}\mu_{ki}^m}, i = \overline{1,c}$$
2) Розрахунок відстаней між новими центрами кластерів і точками даних:
$$D_{ki} = \sqrt{\|X_k - V_i\|^2},$$
$$k = \overline{1,M}, i = \overline{1,c}$$
3) Перерахунок ступенів належності об'єктів кластерам:
$$\mu_{ki} = \frac{1}{\sum_{j=1}^{c}\left(\frac{D_{ki}}{D_{kj}}\right)^{2/m-1}},$$
$$k = \overline{1,M}, i = \overline{1,c}$$

Самоорганізаційна карта Кохонена – нейронна мережа з навчанням без вчителя, що виконує завдання візуалізації і кластеризації [3]. Створює дискретне представлення вхідних просторів навчальних вибірок, які називаються картою. Вихідні сигнали шару Кохонена обробляються за правилом «переможець забирає все»: найбільший сигнал перетворюється в одиничний, інші звертаються в нуль. Самоорганізаційні карти працюють у двох режимах: навчання та відображення. «Навчання» створює карту використовуючи вхідні приклади, тоді як «відображення» автоматично класифікує новий вхідний вектор. Тренування полягає в переміщенні векторів ваги у напрямку зменшення метрики відстані без псування топології. Відмінність самоорганізується карти від звичайної мережі Кохонена полягає в кількості вихідних нейронів: в мережі Кохонена воно повинно відповідати кількості кластерів, а в карті – кількості сегментів, з якого вона повинна складатися, тобто розміром карти. Чим більше число сегментів в карті, тим детальніше вона представляє розподіл об'єктів в просторі ознак. Нейронні мережі та нечітка логіка є лідерами серед методів аналізу даних великого обсягу зі значною кількістю атрибутів.

Адаптивна кластеризація

Для того, щоб кластеризація була більш інформативною і могла застосовуватися до різних завдань виділяють адаптивну методику кластеризації [4]. Під адаптивною кластеризацією розуміється такий аналіз, при якому параметри, що визначають результат, вибираються і коригуються в процесі виконання завдання, виходячи із заданих критеріїв і рекомендацій даних експертом, для досягнення найкращого результату. Адаптивна кластеризація дуже добре підходить до нечітких методів, та нейронних мереж. Відомо багато формул обчислення відстані, кожен з яких придатний для використання в певних випадках (наприклад, квадрат евклідової відстані застосовується для додання більшої ваги більш віддаленим один від одного об'єктів).

Кількість кластерів, порогове значення зупинки роботи алгоритму, спосіб вибору початкових центрів, максимальна кількість ітерацій, кількість одночасно оброблюваних даних, кількість попередніх розділів, коефіцієнт віддаленості, точне значення цих параметрів невідомо і підбирається ітераційним перебором в виділеному інтервалі значень від 2 до |X| кластерів, а також за допомогою експертної оцінки.

Спосіб визначення відстані між кластерами, метод оцінки якості кластеризації, порогове значення для методу оцінки якості кластеризації, початкове граничне значення алгоритму, швидкість навчання мережі вибирає експерт. Використання стандартних значень може привести до дуже поганих результатів. Експертні оцінки параметрів запуску алгоритму будуть усереднюватись і уточнюватись в процесі застосування алгоритму. У загальному вигляді, критерій оцінки якості виконання задачі кластеризації - це чисельний показник, який вираховується за результатами кластеризації на даній ітерації, суть якого – кількісна оцінка якості рішення. Наприклад, показники чіткості досягають максимуму при найбільш чіткому розбитті:

1) Показники чіткості розбиття:
   - Коефіцієнт розбиття:
   $$QR = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K}u_{ij}^2}{Q},$$
   $$QR \in [\frac{1}{K}, 1]$$
   - Індекс чіткості:
   $$CI = \frac{K*PC-1}{K-1},$$
   $$CI \in [0,1]$$

2) Ентропійні критерії (чим меньше значення ентропії, тим краще зроблена кластеризація)
   - Модифікована ентропія:

$$EN = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} u_{qk}\ln u_{qk}}{Q\ln K},$$
$$EN \in [0,1]$$

3) Показник компактності:

$$CP = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} U_{qk}^2 * d^2(x_q, c_k)}{Q * min(d^2(c_i, c_j))},$$
$$i, j \in \overline{1, K}$$

   Мале значення цього критерію каже, що всі наші кластери добре віддільні один від одного, тобто відрізняються, що 1 кластер зберігає 1 клас.

4) Індекс ефективності:

$$PI = \sum_{q=1}^{Q}\sum_{k=1}^{K} u_{qk}^2(d^2(c_k, x) - d^2(x_q, c_k))$$

   Чим більше цей критерій, тим більш оптимальна кількість кластерів у нашій задачі. Ітераційний критерій.

Основна перевага адаптивної кластеризації це можливість визначення кількості кластерів, при якому виходить найкраще розбиття, без залучення експерта. Важливе налаштування адаптивної кластеризації – це вибір критерію оцінки якості рішення. Критерії модифікованої ентропії, ефективності розбиття, модифікований коефіцієнт розбиття – для центроїдної кластеризації, а якість розбиття для нечіткої кластеризації. В якості найкращих рішень вибираються ті розбиття, для яких критерії досягають екстремумів на заданому діапазоні кількості кластерів. Якщо результати задовольняють, то аналіз завершується.

Висновки

- Проведено дослідження існуючих методів і підходів інтелектуального аналізу даних, що використовуються для кластеризації.
- Продемонстровано заходи порівняння схожості об'єктів.
- Дізналися які методи кластеризації більш актуальні для вирішення реальних завдань.
- З'ясовано загальну методику адаптивної кластеризації.
- Було продемонстровано критерії для оцінки якості кластеризації.

Перелік використаних джерел

1. А. Барсегян. Анализ данных и процессов — 3 изд. — 2009. — С. 512 с. — Режим доступа: `http://kist.ntu.edu.ua/textPhD/AnalizDannyhIProcessov.pdf`.
2. Я. Гудфеллоу. Глубокое обучение. — 2018. — 652 с.
3. С. Рашка. Python и машинное обучение. — 2017. — 418 с. — Режим доступа: `https://drive.google.com/file/d/1NOBl5mhAhlTxoxxl8aMWFQMU-MLuZ8ya/view`.
4. O. Wolkenhauer. Fuzzy Clustering. Hard-c-Means, Fuzzy-c-Means, Gustafson-Kessel, Control Systems Centre. — 2001. — 140 с.