УДК 621.391.3

МЕТОД КЛАСТЕРИЗАЦІЇ ДЛЯ ОБРОБКИ ВЕЛИКИХ ОБСЯГІВ ДАНИХ

Ляшенко А.В., Бугаєнко Ю.М.
Інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, Україна
E-mail: andrey.lyashenko44@gmail.com

Clustering method for processing of large data types

The article describes a description of the clustering algorithm for its application in the processing of large volumes of data and in the creation of fuzzy knowledge bases with sets of rules of fuzzy logic.

У статті наводиться опис алгоритму кластеризації для застосування його при обробці великих обсягів даних та при створенні нечітких баз знань з наборами правил нечіткої логіки.

Для великих телекомунікаційних компаній сьогодні постає завдання обробки великих обсягів даних. Потрібно класифікувати дані на різнорідні групи для аналізу даних. Різні види аналізу, для виявлення сучасних тенденцій та збору статистики, потребують математичних обчислень. Щоб вирішити задачу розрізнення об’єктів, використовують кластеризацію. Кластеризація призначена для розділення набору об'єктів на однорідні групи, і її метою є пошук існуючих структур. Цей процес використовується для класифікації трафіку, для побудови маркетингових стратегій оператора зв’язку, крім того в комп'ютерній графіці - для сегментації зображень, для класифікації результатів пошуку, для обробки таблиць і документів. У той же час кожна область даних має свої власні набори даних, наприклад, в системах збору технічних даних необхідно працювати з числовими характеристиками, які мають унікальну оцінку, і, наприклад, при роботі з даними користувача / підприємства дані мають абсолютно інший формат. На основі цього використовуються різні алгоритми кластеризації та обробки даних.

Алгоритми кластеризації оперують з об'єктами. З кожним об'єктом X ототожнюється вектор характеристик $X_i = (x_1, \dots, x_d)$. Компоненти X_i, $i = 1, \dots, d$ є окремими характеристиками об'єкта. Розмірність простору характеристик визначає кількість характеристик d. Що складається з усіх векторів характеристик об'єктів, безліч позначається $M = (X_1, \dots, X_n)$. Підмножина «близьких один до одного» об'єктів з M являє собою кластер. Відстань $D(X_i, X_j)$ між об'єктами X_i і X_j визначається в просторі характеристик на основі обраної метрики.

Часто об’єкти кластеризації мають вимоги до:

- Високої розмірності простору даних - об'єкти описуються великою кількістю атрибутів, отже, повинна бути пристосованість алгоритму до роботи в просторах даних високої розмірності.
- Великий обсяг даних.
- Кластеризація відбувається з метою отримання нечітких правил на основі побудованих кластерів.

Для задоволення цих вимог вибирають метод, який дозволяє кожному об’єкту із вибірки даних знаходитися в кожному кластері із різним ступенем приналежності та має меншу чутливість до викидів. Такий метод має вирішувати задачу визначення належності об’єкта до кластеру, якщо він знаходиться на границі декількох кластерів. До нечітких методів/алгоритмів кластеризації відносять алгоритм нечітких С-середніх *Fuzzy C-means algorithm [3]*.

Алгоритм.

Алгоритм нечітких С-середніх (FCM) схожий на алгоритм К-середніх, так як він порівнює значення об’єкта із вибірки даних із значенням центру кластера. Основна відмінність полягає в тому, що замість того, щоб приймати складне рішення про те, до якого кластеру повинен належати об’єкт, він привласнює значення від 0 до 1, яке описує «наскільки цей об’єкт належить цьому кластеру» для кожного кластера.

Критерії збіжності FCM-алгоритма:

1. Для кожного елемента вибірки вимірювань сума ступенів його приналежності всьому с кластерам повинна дорівнювати

$$\sum_{i=1}^{c} \mu_{ij} = 1, \qquad \forall j = 1, \ldots . N,$$

2. Значення ступеня приналежності повинно бути обмежено інтервалом [0,1]:

$$\mu_{ij} \in [0,1], \forall i = 1, \ldots, c, \text{ і } \forall j = 1, \ldots . N.$$

Нечітке правило говорить, що сума значення членства об’єкта для всіх кластерів повинна дорівнювати 1. Чим вище значення членства, тим більш імовірно, що об’єкт належить цьому кластеру. Кластеризація FCM здійснюється шляхом мінімізації цільової функції, показаної в рівнянні (1):

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{q} \ |x_i - v_k|^2 \ (1),$$

Де:

J – цільова функція;

n – кількість об’єктів у виборці даних;

c – кількість кластерів;

μ – нечітке значення членства з таблиці;

q – коефіцієнт нечіткості (значення > 1);

x_i – значення і-ого об’єкта із виборки;

v_k – центр кластера ;

$|x_i - v_k|$ – Евклідова відстань, яка визначається рівнянням (2):

$$|x_i - v_k| = \sqrt{\sum_{i=1}^{n} (x_i - v_k)^2}$$

Розрахунок центра кластера визначається за допомогою рівняння (3):

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{q} x_i}{\sum_{i=1}^{n} \mu_{ik}^{q}}$$

Таблиця нечіткого членства розраховується за допомогою рівняння (4):

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{|x_i - v_k|}{|x_i - v_l|}\right)^{\frac{2}{q-1}}}$$

Кроки реалізації:

Крок 1: Встановити кількість кластерів та нечіткий параметр (константне значення > 1) і параметр зупинки

Крок 2: Ініціалізація матриці степенів приналежності

Крок 3: Встановити лічильник циклів $k = 0$

Крок 4: Обчислити центроїди кластера, обчислити значення цільової функції J

Крок 5: Для кожного об'єкта та для кожного кластера обчислити значення членства в матриці

Крок 6: Якщо значення J між послідовними ітераціями менше, ніж умова зупинки, то зупинка; інакше встановіть $k = k + 1$ та перейти до кроку 4.

Крок 7: Отримання матриці приналежності та кінець алгоритму

Виходячи з цього алгоритму можна зробити висновок, що такий підхід дозволяє визначити належність об'єкта з вибірки даних до кластерів із різним ступенем приналежності. Це дозволяє не втрачати наявні логічні зв'язки на границі кластерів.

Висновки. В даній статті був запропонований покращений спосіб обробки даних, який базується на алгоритмі кластеризації, який має назву алгоритм нечітких С-середніх (*Fuzzy C-means algorithm).* Цей алгоритм дозволяє належати кожному об'єкту з вибірки даних одразу декільком кластерам, що дозволяє ще більш точно(«м'яко») розділити об'єкти, які мають велику розмірність простору даних (багато атрибутів) і також великий обсяг даних, між різними групами, в залежності від предметної області кластеризації.

Література

1. Approach to determining the number of clusters in a data set. Ivan Ishchenko, Larysa Globa, Yurii Buhaienko, Andrii Liashenko National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" Kyiv.
2. Нечеткое моделирование и управление. А. Пегат с. [520-530].
3. Fuzzy C-means Algorithms for very large data. Timothy C. Havens, James C. Bezdek, Cristopher Leckie, Lawrence O. Hall, Marimuthu Palaniswami. IEEE Transactions on Fuzzy Systems 2012.