УДК 004.9
Марта Дубик
Західноукраїнський національний університет

ПІДВИЩЕННЯ ТОЧНОСТІ КЛАСТЕРИЗАЦІЇ ВЕЛИКИХ ДАНИХ НА ОСНОВІ НЕЙРОМЕРЕЖЕВИХ МОДЕЛЕЙ

Marta Dubyk

IMPROVING THE ACCURACY OF CLUSTERING LARGE DATA BASED ON NEURAL NETWORK MODELS

З розвитком великих даних і збільшенням їхньої різноманітності та складності, традиційні методи кластеризації часто виявляються недостатньо ефективними. Спектральна кластеризація, яка базується на використанні власних векторів графів подібності, дозволяє виявити складні структури в даних, але вона також має обмеження при роботі з великими обсягами даних. З іншого боку, сіамські нейронні мережі, які ефективно використовуються для визначення подібності між об'єктами, можуть бути застосовані для поліпшення процесу кластеризації.

Розглянемо узагальнений алгоритм підвищення точності кластеризації за допомогою сіамських нейронних мереж і спектральної кластеризації:

1. Попередня обробка даних:
   - виконати нормалізацію та стандартизацію даних;
   - застосування методів зниження розмірності, якщо необхідно.
2. Застосування сіамської нейронної мережі для визначення подібності між об'єктами:
   - навчити сіамську нейронну мережу на основі пар схожих та несхожих об'єктів з датасету для отримання простору ознак, який відображає подібність між об'єктами;
   - використати отримані вектори ознак для створення матриці подібності.
3. Виконання спектральної кластеризації:
   - побудувати граф подібності на основі матриці, отриманої з сіамської нейронної мережі;
   - застосувати спектральну кластеризацію до графу, використовуючи власні вектори для розподілу об'єктів на кластери.
4. Оптимізація та налаштування:
   - підлаштувати параметри сіамської нейронної мережі та спектральної кластеризації для досягнення оптимальної точності;
   - використовувати крос-валідацію для перевірки стійкості моделі.
5. Оцінка результатів: оцінити якість кластеризації за допомогою метрик.

Цей підхід дозволяє ефективно об'єднати переваги сіамських нейронних мереж у визначенні складних відносин між об'єктами з можливостями спектральної кластеризації для розпізнавання тонких групових структур. Це особливо важливо при роботі з великими наборами даних, де традиційні методи можуть не виявляти складних взаємозв'язків або ж вимагають неприпустимо високих обчислювальних ресурсів.

Література

1. Wang C., Shakhovska N., Sachenko A., Komar M. A New Approach for Missing Data Imputation in Big Data Interface. Information Technology and Control. 2020. Vol. 49. No 4. Pp. 541-555.

2. Комар М.П. Методи відновлення відсутніх даних у інтерфейсі великих даних. Вимірювальна та обчислювальна техніка в технологічних процесах. 2020. №5. С. 97–103.