

АНАЛІЗ ЗАСОБІВ КЛАСТЕРИЗАЦІЇ ДАНИХ

Кузьменко О.В., Черних О.П.

Національний технічний університет

«Харківський політехнічний інститут», Харків, Україна

На сьогодні зростання обсягів даних з різними властивостями приводить до необхідності аналізу швидкодіючих і надійних засобів кластеризації.

Рішення завдання кластеризації принципово неоднозначно за декількома причинами: досить немає однозначно найкращого критерію якості кластеризації та число кластерів, як правило, невідомо заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію. Важко забезпечити чітку категоризацію методів кластеризації, оскільки вони можуть перекриватися, так що метод може мати функції з декількох категорій. Методи кластеризації можна розділити на два основних типа: ієрархічна і секційна кластеризації.

У загальному вигляді кластерний аналіз включає наступні етапи: відбір вибірки об'єктів для кластеризації; визначення безлічі змінних, за якими будуть оцінюватися об'єкти у вибірці; обчислення значень міри схожості між об'єктами; застосування методу кластерного аналізу для створення груп схожих об'єктів (кластерів); представлення результатів аналізу. Визначено, що загальна схема кластеризації одна, але існує багато реалізацій цієї схеми.

Однією з цілей кластеризації є виявлення внутрішніх зв'язків між даними шляхом визначення кластерної структури. Розбиття спостережень на групи схожих об'єктів дозволяє спростити подальшу обробку даних і прийняття рішень, застосовуючи до кожного кластеру свій метод аналізу. Після отримання та аналізу результатів можливе корегування обраного методу кластеризації до отримання оптимального результату.

Для визначення оптимального числа кластерів в роботі були розглянуті алгоритм пов'язаних компонентів і алгоритм пошарової кластеризації. За допомогою пошарової кластеризації були обрані пов'язані компоненти на різних рівнях і задана потрібна глибина одержуваних кластерів, що, в свою чергу, прискорило роботу кластерів.

Однак, при загальному використанні будь-якого алгоритму важливо розуміти його достоїнства і недоліки та обов'язково враховувати природу даних, з якими він краще працює.

Список літератури

1. Олдендерфер М.С., Блэшфилд Р.К. Кластерный анализ / Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989 — 215 с.
2. Шлезингер М. Десять лекций по статистическому и структурному распознаванию. // Шлезингер М., Главач В. — Киев: Наукова думка, 2004. — 343 с.