УДК 519.237.8

Н.Е. КОНДРУК, М.М. МАЛЯР

АНАЛІЗ КЛАСТЕРНИХ СТРУКТУР ЗА РІЗНИМИ МІРАМИ ПОДІБНОСТІ

Анотація. Наведено аналіз кластерних утворень, що використовують в практичних задачах. У різних дослідженнях сегментацію даних зазвичай виконують лише одною формою кластерів. Запропоновано здійснювати кластеризацію за різними мірами подібності одних і тих самих досліджуваних даних та виявляти різні види взаємозв'язків між ними. Це дає змогу проводити більш повний, різнобічний та системний аналіз утворених сегментів у прикладних задачах. Верифікацію цього підходу реалізовано на практичній задачі аналізу демографічних процесів у низці європейських країн.

Ключові слова: кластеризація, аналіз кластерів, інтерпретація кластерів, демографічні процеси.

ВСТУП

Останнім часом технологію кластеризації використовують для аналізу даних різної природи, зокрема за відсутності попередньої інформації про кількість кластерів та навчальну вибірку. Спектр застосування кластерного аналізу дуже широкий та представлений в багатьох роботах [1–12]. Це дослідження присвячене підходам до кластеризації, за допомогою яких потрібно визначити не лише кластерну структуру об'єктів (як у задачах розпізнавання образів [1]), а й змістовно інтерпретувати отримані сегменти (наприклад, ринку, цільової аудиторії, у маркетингових дослідженнях тощо) [2–4]. При цьому обґрунтованість висновків інколи не є достатньо переконливою, оскільки можливо сформувати різні види кластерів на основі тих самих даних залежно від методу або способу проведення аналізу. Важливе значення для отримання коректних результатів має вибір міри подібності, яка не спотворює взаємозв'язків між об'єктами у разі, якщо ці взаємозв'язки потребують змістовної інтерпретації [1]. Це є першим кроком до визначення валідності кластерів, який відбувається ще до етапу аналізу. В одному методі кластеризації реалізують тільки одну міру подібності. У багатьох програмних пакетах і алгоритмах зазвичай використовують евклідову відстань, що зумовлює утворення еліпсоїдних кластерів. Тому пропонується проводити аналіз даних за різними мірами подібності. Придатним інструментарієм для цього є метод кластеризації, що базується на нечітких бінарних відношеннях [5] та здійснює кластеризацію еліпсоїдну, конусну та концентричними сферами.

ОГЛЯД СУЧАСНИХ ДОСЛІДЖЕНЬ З ПРИКЛАДНОЇ КЛАСТЕРИЗАЦІЇ

Інтерпретація кластерної структури наборів даних є одним із важливих етапів прикладної кластеризації. Метою проведення кластерного аналізу в таких задачах є аналіз утвореної сегментації. Це дає можливість отримувати нову необхідну інформацію для прийняття рішень у багатьох сферах діяльності.

Зокрема, в [6, 7] здійснено кластерний аналіз поведінки користувачів електроенергії під час використання розумних лічильників для планування роботи розподільних мереж. Було сформовано десять різних груп поведінки клієнтів. У праці [8] визначено групи ризику набуття надмірної ваги та виникнення ожиріння серед дітей і підлітків на основі показників фізичної активності та харчових звичок. Поведінку користувачів за даними потоків кліків у реальних соціальних мережах