УДК 519.237.8
Н.Е. Кондрук
Кандидат технічних наук, доцент, доцент кафедри кібернетики і прикладної математики
ДВНЗ «УжНУ», Ужгород

ВИКОРИСТАННЯ РІЗНИХ ВИДІВ МІР ПОДІБНОСТІ В КЛАСТЕРНОМУ АНАЛІЗІ

Вступ. Кластеризація (неконтрольована класифікація, автоматичне групування об'єктів) дозволяє розбити набір даних на деяку кількість однорідних в певному сенсі кластерів. Центроїдна неконтрольована класифікація реалізується методами кластерного аналізу і дозволяє виявляти властивість даних групуватись близько до деяких значень (центрів). Загалом концепцію кластерного аналізу багатовимірних даних можна визначити як розподіл всіх можливих точок (об'єктів) m-вимірного простору ознак за відповідними кластерами. При цьому об'єкти одного кластера групуються в просторі ознак компактно: подібність об'єктів всередині кластера більша, ніж між об'єктами різних кластерів. Поняття подібності об'єктів математично може бути виражено різними способами. Найчастіше для цього використовується деяка метрика відстані (Евкліда, манхеттенська, Чебишева та ін.). При цьому форма кластерів обмежується еліпсоїдною. Існує цілий ряд прикладних задач розв'язання яких потребує утворення кластерів інших геометричних форм [1-3].

Таким чином, доцільним є розробка математичного апарату, який дозволив би проводити групування об'єктів на кластери різних геометричних форм. Це дає можливість ефективно розв'язувати достатньо широкі класи прикладних задач із різних предметних областей.

Деякі види мір подібності. Пропонується використати міру подібності експоненціального виду засновану на поняті нечітких бінарних відношень об'єктів [2, 4]:

$$\mu_R(\overline{c_i},\overline{c_j}) = e^{-\rho_{ij}}, \qquad (1)$$

де R - нечітке бінарне відношення задане на множині векторних ознак $C = \{\overline{c_i} | i = \overline{1,m}\}$ із функцією належності $\mu_R : C^2 \to [0,1]$, що характеризує подібність об'єктів за деяким критерієм, $\rho_{ij} : C^2 \to [0,1]$ - деякий функціонал. Вибір саме такого виду функції належності забезпечує можливість підбору ρ_{ij} так, щоб близькість її значень до 1 характеризувала сильну подібність, а до 0 відмінність об'єктів i та j за певним критерієм. Функція виду (1) має «хороші» властивості – гладкість, неперервність та монотонність.

В роботі [4] розроблено алгоритм однорівневої кластеризації, що використовує, зокрема, міри подібності об'єктів типу (1) - R^V, R^D. Ідея методу полягає у тому, щоб задати подібності за певним нечітким бінарним відношенням і утворенням кластеру, із тих об'єктів, функція належності μ_R яких більша за

певний поріг кластеризації – число із проміжку [0, 1]. Проведення практичних експериментів показало, що «хороша» чутливість функції типу (1) в околі свого граничного значення ($\sup \mu_{R^D}=1$) дозволяє проводити кластеризацію об'єктів для всіх можливих величин порогів проміжку [0; 1] із певною точністю (наприклад, із точністю 0,01). Це забезпечує можливість проводити дослідження всієї динаміки зміни кластерів та їх структури.

Так нечітке бінарне відношенням R^V [4] характеризує близькість точок простору ознак об'єктів за відстанню і приводить до утворення еліптичних кластерів. «Довжинна» міра подібності R^D дозволяє розбивати вектори ознак об'єктів на кластери концентричними сферами [2] і характеризує різницю довжин векторів ознак обєктів.

Кластеризація еліптичними кластерами є найбільш поширена при розв'язані багатьох прикладних задач, так як схожість об'єктів проводиться на основі «відстаневої» міри подібності. Але використання саме однорівневого методу показало дуже хороші результати на практиці, які описані в [4]. Кластеризація концентричними кластерами (кластерами у формі концентричних сфер) [2] дала можливість групувати об'єкти за довжинною схожістю їх векторів ознак та отримувати якісно нову прикладну змістовну інтерпретацію утворених однорідних груп на практиці.

Висновок. Отже, проведені дослідження показали, що міри подібності, засновані на нечітких бінарних відношеннях, які характеризуються функцією належності типу (1) забезпечують проведення групування точок простору ознак кластерами різних геометричних форм і можуть бути ефективно використані для розв'язання різних прикладних задач кластерного аналізу.

Список використаних джерел

1. Кондрук Н. Е. Системи підтримки прийняття рішень для автоматизованого складання дієт / Н. Е. Кондрук // Управління розвитком складних систем. – 2015. – Вип. 23(1). – С. 110–114.

2. Кондрук, Н. Е. Використання довжинної міри подібності в задачах кластеризації / Н. Е. Кондрук // Радіоелектроніка, інформатика, управління. – 2018. – № 3 (46) – С. 98-105. DOI: 10.15588/1607-3274-2018-3-11.

3. Кондрук, Н. Е. Деякі методи автоматичного групування об'єктів / Н. Е. Кондрук // Південно-Європейский журнал передових технологій. – 2014. – Т. 2. – № 4 (68). – С. 20–24.

4. Kondruk N. Clustering method based on fuzzy binary relation / N. Kondruk // Eastern-European Journal of Enterprise Technologies. - 2017. - № 2(4). - С. 10-16. DOI: 10.15587/1729–4061.2017.94961 2.