

УДК 681.518:004.93.1

КЛАСТЕРИЗАЦІЯ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ ПРИ ДИСТАНЦІЙНОМУ НАВЧАННІ

С.О. Петров, асистент СумДУ

При синтезі адаптивних систем керування дистанційним навчанням (СКДН), на етапі формування початкової конфігурації розбиття простору ознак на класи розпізнавання доцільним є застосування алгоритмів кластер-аналізу, що дозволяє:

- автоматизувати процес формування навчальної матриці;
- підвищити ефективність функціонування СКДН, а відповідно й ефективність навчального процесу.

Існуючі методи кластеризації поділяють на дві групи: декомпозиційні, та ієрархічні. Декомпозиційні (k -кластеризація) – це методи, які базуються на одностов'язності об'єктів і контейнерів розпізнавання, ієрархічні – працюють в тих випадках якщо деяка достатньо велика група об'єктів включає в себе групу меншого розміру.

Особливість відомих алгоритмів ієрархічної класифікації (Single Link, Complete Link, Group Average) полягає в тому, що вони розбивають вектори реалізації на кластери, шляхом розбиття їх на ієрархічні групи, що дозволяє зменшити вплив багатовимірності. Такі методи кластеризації ще мають назву агломеративні, в загальному випадку, використовують махалонобіусову метрику, яка в частковому випадку має вигляд:

$$d_{ij}^2 = \sum_{k=1}^q v_k (z_j^{(k)} - z_i^{(k)})^2 \quad (1)$$

де $z_i^k = (U'_k, X_i)$ – одновимірна проекція лінійних комбінацій вхідних змінних; $v_k = \varphi(Q_k)$ – лінійна

комбінація $U_i, \forall i$ в новому базисі коваріаційної матриці ковариационной матриці S ; $\varphi(\cdot)$ - деяка монотонно-зростаюча функція. Ідея алгоритмів декомпозиційної кластеризації (K-means) ідея яких полягає в представленні кластера у вигляді центроїда, який є центром мас усіх векторів що входять в кластер. Тут вхідними величинами є матриця S та число k , вводиться деяка оціночна функція $c: \{X : X \leq S\} \rightarrow R^+$ - вартість кластера, потім ставиться оптимізаційна задача вибору такої множини векторів, яка буде мінімізувати вартість кластера.

$$\min \sum_i c(S_i), i = 1 \dots k \quad (2)$$

Основними недоліками існуючих методів кластеризації є відсутність загального критерію оцінки оптимальності побудови початкової конфігурації класів та немає загальних рекомендацій по апріорному визначенню числа k .

Метод, який розроблений автором базується на гібридному алгоритмі, в якому комбінуються критерії (1) і (2), та проводиться обчислення ентропійного критерію Шенона, як показника якості розбиття, який обчислюється в рамках інформаційно-екстремальної технології за методом функціонально статистичних випробувань. Цей метод дозволяє з заданою ефективністю розв'язувати задачу кластер-аналізу, маючи загальний критерій оцінки якості розбиття та не потребуючи апріорно визначеного значення числа k . Розроблено практичну реалізацію алгоритму, та проведено експеримент по кластеризації даних які отримані з відповідей студентів на тестові запитання. За результатами тестування було сформовано навчальну матрицю, яка була кластеризована на чотири класи, які відповідають оцінкам "відмінно", "добре", "задовільно", "незадовільно".