УДК 629.01

ЗАСТОСУВАННЯ МЕТОДУ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ К-СЕРЕДНІХ В ЗАДАЧАХ РОЗПІЗНАВАННЯ СТАНУ ПАЦІЄНТІВ В СИСТЕМАХ МЕДИЧНОГО МОНІТОРИНГУ

Заярна Вікторія Дмитрівна, студентка группы 355
Національний аерокосмічний університет шм. М.Е. Жуковського «ХАИ»

Математичний апарат кластеризації широко застосовується в діагностичних цілях, розв'язанні класифікаційних завдань та пошуку нових закономірностей, для встановлення нових наукових гіпотез. У даній роботі розглядається актуальне питання кластеризації даних у медицині.

Було проведено аналіз основних методів кластеризації, а також обґрунтування вибору методу к-середніх. Його основними перевагами є універсальність, швидкість і простота програмної реалізації. Також метод к-середніх гнучкий до використання різноманітних метрик та змін.

Алгоритм k-середніх будує k кластерів, розташованих на можливо великих відстанях один від одного. Основний тип задач, які вирішує алгоритм k-середніх – наявність припущень (гіпотез) щодо кількості кластерів, при цьому вони повинні бути різні настільки, наскільки це можливо. Вибір кількості k може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

У даній роботі застосовується метод нечіткої кластеризації для модифікації методу к-середніх, що дозволяє кожному об'єкту належати з різною мірою кільком або всім кластерам одночасно. Число кластерів вважається заздалегідь відомим.

Початковою інформацією є вибірка спостережень, сформована з N n-мірних векторів ознак $X = \{x(1), x(2),..., x(N)\}$, $x(k) \in X$, $k=1,2,...,N$. Результат роботи методу є розбиття початкового масиву даних на m классов класів з деяким рівнем $w_j(k)$ належності k-того вектора ознак j-му кластеру. Цільова функція, що підлягає мінімізації має вигляд:

$$E(w_j(k), c_j) = \sum_{k=1}^{N} \sum_{j=1}^{m} w_j^{\beta}(k) d^2(x(k), c_j) \to min,$$

при обмеженнях:

$$\sum_{j=1}^{m} w_j(k) = 1, k = 1, \dots, n, 0 < \sum_{k=1}^{N} w_j(k) < N, j = 1, \dots, m.$$

Тут w_j (k)$\in$[0,1] – рівень належності вектора $x(k)$ до j-го кластера, c_j – центроїд j-го кластера, $d^2(x(k), c_j)$– відстань між $x(k)$ та c_j в прийнятій метриці, β – невід'ємні параметр, що іменується «фаззіфікатором» (в разі використання $d^2(x(k), c_j)$ в якості евклідової відстані, приймається рівним 2).

Робота алгоритму починається з завдання початкової випадкової матриці нечіткого розбиття W_0. Відповідно до її значеннями вираховується початковий набір центрів прототипів c_j^0, згідно з формулою

$$c_j = \frac{\sum_{k=1}^{N} w_j^{\beta}(k)x(k)}{\sum_{k=1}^{N} w_j^{\beta}(k)}.$$

На підставі розрахованих центрів-прототипів c_j^0 далі обчислюється матриця W_1 згідно з формулою

$$w_j = \frac{(d^2(x(k), c_j))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}(d^2(x(k), c_l))^{\frac{1}{1-\beta}}}.$$

Після цього в пакетному режимі перераховуються $c_j^1, W^2, \dots, W^t, c_j^t, W^{t+1}$ і так далі до тих пір, поки різниця між нинішніми і наступними значеннями матриці W не стане менше заданого порогу точності. Таким чином, вся наявна вибірка даних обробляється багаторазово.

В результаті роботи алгоритму отримаємо матрицю нечіткого розбиття, в якій пацієнти будуть розділені на кластери (діагнози). Форма кластерів може змінюватися від гіпершара до гіперелліпсоіда в залежності від форми вихідних даних, тобто від вибору відстані між *x(k)* та c_j:

$$d(x(k), c_j) = \sqrt{(x(k) - c_j)^T A_j (x(k) - c_j)},$$

де A_j - матриця, яка може бути визначена як зворотна нечітка ковариационная матриця кожного кластера.

Якщо в якості матриці *Aj* візьмемо одиничну матрицю, то в результаті отримаємо евклідову відстань $d(x(k), c_j) = \sqrt{(x(k) - c_j)^T (x(k) - c_j)}$, і форма кластерів буде округла (гіпершари).

Для додання кластерам форми гіперелліпсоідов як матрицю A_j можна використовувати симетричну позитивно визначену матрицю, тобто матрицю, у якій всі власні значення є дійсними і позитивними і $A_j = F_j^{-1}$, де

$$F_j = \frac{\sum_{j=1}^{m} w_j^{\beta}(k)(x(k)-c_j)(x(k)-c_j)^T}{\sum_{j=1}^{m} w_j^{\beta}(k)}.$$

В результаті роботи алгоритму кластеризації ми отримуємо поділ наших даних на однорідні кластери, які можуть мати форму довільно орієнтованих в просторі гіперелліпсоідов. Також буде відома ступінь належності кожного з об'єктів до кожного з кластерів $w_j(k)$.

** Науковий керівник – Бакуменко Н.С., к.т.н., доцент каф. 304.*