Rozmyta mediana i sposoby jej wyznaczania

Dla ostrego zbioru danych X_h medianą nazywamy wartość środkową w uporządkowanym niemalejąco zbiorze:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}, \tag{1}$$

gdy n jest nieparzyste, bądź średnią arytmetyczną dwóch środkowych wartości, gdy n jest parzyste, tzn.

$$m_h = \begin{cases} x_{\frac{n}{2}} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \end{cases}. \tag{2}$$

Aby zdefiniować zbiór rozmyty X_f oprócz wartości elementów x_k wymagana jest znajomość ich stopnia przynależności u_k. Rozmytą medianą m_f zbioru X_f jest taka wartość, że suma stopni przynależności na lewo wraz z medianą oraz na prawo wraz z medianą jest taka sama. Wartość rozmytej mediany jak i mediany ostrej nie jest jednoznaczna. W przypadku dużej ilości danych (n rzędu 16K bądź więcej) uporządkowanie ich staje się nieopłacalne ze względu na wielkość zajmowanej przestrzeni oraz wymagany czas na ich uporządkowanie. Innym sposobem wyznaczania mediany jest wykorzystanie definicji bazującej na funkcjonale. Ostrą medianą zbioru danych X jest wartość m_h minimalizująca funkcjonał:

$$P_h(m) = \sum_{k=1}^{n} \left| x_k - m \right|. \tag{3}$$

W przypadku zbiorów rozmytych mediana minimalizuje funkcjonał postaci:

$$P_f(m_f) = \sum_{k=1}^{n} u_k \left| x_k - m_f \right|. \tag{4}$$

W praktyce wyznaczenie mediany sprowadza się do znalezienia pierwiastka pochodnej funkcjonału (4) w przypadku ostrej mediany:

$$\Psi_h(m_h) = \sum_{k=1}^{n} \operatorname{sgn}(x_k - m_h) = 0, \tag{5}$$

zaś dla mediany rozmytej:

$$\Psi_f(m_f) = \sum_{k=1}^{n} u_k \operatorname{sgn}(x_k - m_f) = 0. \tag{6}$$

Modyfikacja metody wyznaczania rozmytej mediany

Rozwiązując równanie (6) dla zbioru rozmytego otrzymujemy wartość rozmytej mediany. We wzorze tym występuje funkcja:

$$\operatorname{sgn}(d) = \begin{cases} -1, \text{dla } d < 0 \\ 0, \text{dla } d = 0 \\ 1, \text{dla } d > 0 \end{cases} \tag{7}$$

Funkcja ta jest nieciągła w d=0 – uniemożliwia to zastosowanie innych metod poszukiwania pierwiastka niż metoda bisekcji. W literaturze [1] spotyka się rozmytą funkcję signum o postaci:

$$\operatorname{sgn}_f(x) = \frac{x}{\sqrt{x^2 + d^2}} \tag{8}$$

gdzie d-parametr odpowiedzialny za „zmiękczenie” działania funkcji.

Zmieniając postać równania (6) poprzez zastąpienie ostrej funkcji signum funkcją rozmytą otrzymujemy równanie, którego rozwiązaniem jest estymata rozmytej mediany zbioru rozmytego X.

$$\hat{\Psi}_f(m_f) = \sum_{k=1}^{n} u_k \operatorname{sgn}_f(x_k - m_f) = 0 \tag{9}$$

Funkcja $\hat{\Psi}_f$ jest funkcją ciągłą w związku z czym do wyznaczenia jej pierwiastka zastosowano metodę regula-falsi [2].

Bibliografia

[1] P. Kersten, “Implementation Issues in the Fuzzy c-Medians Clustering Algorithm”, Fuzz-IEEE’97, pp. 957-962

[2] W.H. Press, S.A. Teukolsky, ...,”Numerical Recipes in C”, Cambridge Univ. Press, NY USA 1988 1992.