Granular Computing 9999
pages 1–4

METODY SZTUCZNEJ INTELIGENCJI - PROJEKTY

PB

1 Projekt z grupowania danych - Fuzzy k-means Clustering

Liczba osób realizujących projekt: 1 osoba

1. Wczytanie danych w formatach arff, tab
2. Wybór atrybutów, które mają zostać uwzględnione podczas grupowania
3. Pobranie parametrów algorytmu k-średnich, w tym:
   (a) współczynnik rozmytości
   (b) liczba iteracji, ewentualnie brak zmian w wynikowych środkach klas
   (c) liczba grup (skupień, klas)
4. Wypisanie wyników grupowania, przydzielenie do poszczególnych grup
5. Zapisanie wyniku pogrupowania z dodaniem jednego atrybutu (kolumny) określającej numer grupy poszczególnych obiektów (format arff, tab).

1.1 Nazewnictwo

$(x_1, x_2, ...)$ - zbiór obiektów, reprezentujących dane
$x_i = \{x_i^1, x_i^2, .., x_i^p\}$, gdzie x_i^j oznacza atrybut o indeksie j obiektu x_i.
U przestrzeń wszystkich obiektów
X - podzbiór zbioru wszystkich obiektów U
x_i - obiekt należący do podzbioru wszystkich obiektów U
A - zbiór wszystkich atrybutów, cech, właściwości
a_i - atrybut należący do zbioru atrybutów A
V_{a_i} - zbiór wszystkich wartości atrybutu a_i (nazywany dziedziną a_i)
$V(a_i)$ - zbiór wszystkich wartości atrybutu a_i (nazywany dziedziną a_i)
B - niepusty podzbiór A ($B \subseteq A$)
$LOW(X_B)$ - dolna aproksymacja X względem B
$\underline{X}_B$ - dolna aproksymacja X względem B
$UPP(X_B)$ - górna aproksymacja X względem B
$\overline{X}_B$ - górna aproksymacja X względem B
AS_B - standardowa przestrzeń aproksymacyjna
$AS_{\#,\$}$ - sparametryzowana przestrzeń aproksymacyjna
$R_{a_i}(X)$ - przybliżoność ze wzdgledu na $\{a_i\}$
$Rough_{a_j}(a_i)$ - średnia przybliżoność atrybutu a_i względem atrybutu $\{a_j\}$
$MR(a_i)$ - minimalna przybliżoność atrybutu a_i
MMR - minimalna wartość MR wszystkich atrybutów
$IND(B)$ - relacja nierozróżnialności
$[x_i]_{IND(B)}$ - klasa równoważności obiektu x_i w relacji $IND(B)$, nazywana także zbiorem elementarnym w B
$(C_1, C_2, \ldots, C_K)$ - klasy, skupienia w danym pogrupowaniu danych
$Card(X)$ - liczebność zbioru X
$|X|$ - liczebność zbioru X
$P(U)$ - zbiór potęgowy zbioru U

2 Fuzzy k-Means Clustering

2.1 Wprowadzenie

2.2 Algorytm

Niech $U = \{x_1, x_2, ..., x_n\}$ będzie przestrzenią n obiektów, a $C = \{v_1, v_2, ..., v_c\}$ będzie zbiorem c centroidów, gdzie $x_j \in \mathbf{R}^m$, $c_i \in \mathbf{R}^m$. Algorytm fuzzy k-means stanowi wersję operującą na zbiorach rozmytych algorytmu k-średnich. Dokonuje podziału przestzreni U na c klas poprzez minimalizowanie następującej funkcji celu:

$$J = \sum_{j-1}^{n} \sum_{i-1}^{c} (\mu_{ij})^{\acute{m}_1} ||x_j - v_i||^2 \tag{1}$$

gdzie $1 \leq \acute{m}_1 \leq \infty$ jest współczynikiem określającym rozmytość, v_i jest i-tym centroidem, odpowiadającym klasie C_i, $\mu_{ij} \in [0, 1]$ jest funkcją przynależności obiektu x_j do klasy C_k, $||.||$ jest normą, w ten sposób

$$v_i = \frac{1}{n} \sum_{j-1}^{n} (\mu_{ij})^{\acute{m}_1} x_j \tag{2}$$

$$\mu_{ij} = \left\{ \sum_{k-1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{\acute{m}_1 - 1}} \right\}^{-1} \tag{3}$$

pod warunkiem, że

$$\sum_{i-1}^{c} \mu_{ij} = 1, \forall j, \ 0 < \sum_{j-1}^{n} \mu_{ij} < n, \forall i \tag{4}$$

Algorytm rozpoczyna działanie poprzez przydzielenie losowo c środków klas jako centroidów (średnich) dla c klas. Wartości przynależności obiektów do klas zostają obliczone na podstawie odległości obiektu od środka klasy $\{v_i\}$ weług równania 4. Po określeniu przynależności wszystkich obiektów przestrzeni, nowe środki klas zostają obliczone według równania 3. Algorytm kończy działanie w momencie ustabilizowania się środków klas w kolejnych iteracjach. Praktycznie oznacza to warunek, aby środki klas z poprzedniej iteracji pokrywały się z środkami klas kolejnej iteracji. Podstawowe kroki algorytmu podano w tabeli 1.

Algorithm 1: Fuzzy k-Means Clustering

Data: Input Data
Result: Fuzzy k-means

1) Przydziel losowo początkowe środki klas $v_i, i = 1, 2, ...c$. Wybierz wartości $\acute{m}_1$ oraz próg ϵ. Licznik iteracji ustaw na 1.

2) Oblicz wartości przynależności μ_{ij} obiektów do klas według równania 3 dla c klas i n obiektów.

3) Przelicz na nowo środki klas v_i według równania **??**

4) Powtarzaj kroki od 2 do 4 zwiększając licznik t, dopóki $|\mu_{ij} - \mu_{ij}(t-1)| > \epsilon$