Granular Computing 9999
pages 1–3

METODY SZTUCZNEJ INTELIGENCJI - PROJEKTY

PB

1 Projekt z grupowania danych - Rough k-means

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 Rough *k*-means Clustering

Algorytm:

Zbiór danych: X_n - n-ty punkt danych oraz $X = (X_1, ..., X_n)^T$
C_k - k-ta grupa (skupienie, klasa)
$\underline{C_k}$ - aproksymacja dolna klasy C_k
$\overline{C_k}$ - aproksymacja górna klasy C_k
$C_k^B = \overline{C_k} - \underline{C_k}$ - brzeg klasy C_k
m_k - średnia klasy C_k
$M = (M_1, ..., M_k^T$ z $k = 1, .., K$
Odległość między punktem X_n a środkiem m_k: $d(X_n, m_k) = ||X_n - m_k||$

Krok 0: Inicjalizacja

Losowo przydziel każdy punkt danych do jednej i tylko jednej aproksymacji dolnej i odpowiadającej jej aproksymacji górnej.

Krok I: Obliczenie nowych średnich

$$m_k = \begin{array}{c} w_l * \Sigma_{X_n \in \underline{C_k}} \frac{X_n}{|\underline{C_k}|} + w_b * \Sigma_{X_n \in C_k^B} \frac{X_n}{|C_k^B|} \quad dla \quad C_k^B \neq 0 \\ w_l * \Sigma_{X_n \in \underline{C_k}} \frac{X_n}{|\underline{C_k}|} \quad w\ przeciwnym\ razie \end{array}$$

gdzie w_l oraz w_b oznaczają wagi. Symbol $|\underline{C_k}|$ oznacza liczbę obiektów (punktów) w dolnej aprosymacji, $|C_k^B|$ oznacza liczbę obiektów brzegu klasy $|C_k^B| = |\overline{C_k} - \underline{C_k}|$.

Krok II: Przydziel obiekty do aproksymacji

(1) Dla danego obiektu X_n, znajdź najbliższy środek klasy m_k:

$d_{n,h}^{min} = d(X_n, m_h) = \min_{k=1,...,K} d(X_n, m_k)$

Przydziel X_n do górnej aproksymacji klasy h: $X_n \in |\overline{C_h}|$

(2) Określ zbiór T dla danej wartości progowej ϵ:

$$T = t : d(X_n - m_h) - d(X_n - m_k) \leq \epsilon \wedge h \neq k.$$

$$IF(T \neq 0)\ Then\ \{X_n \in \overline{C_k}, \forall t \in T\}\ Else\ \{X_n \in \underline{C_k}\}$$

Krok III: Sprawdź zbieżność algorytmu

$If(Algorytm\ spelnia\ warunek\ stopu)\ Then\ \{STOP\}\ Else\ \{Przejdz\ do\ kroku\ I\}.$