grupowanie danych niejednorodnych

[NTwI] Obliczenia ziarniste

projekt, 2024

Problem

Nieraz dane, którymi dysponujemy, nie są jednorodne. Przykładowo pochodzą z różnych źródeł i nie mają takich samych atrybutów. W przypadku brakujących wartości atrybutów można je próbować uzupełniać. Metod uzupełniania jest wiele. Jedną z nich jest uzupełnianie na podstawie wartości k najbliższych sąsiadów. Jednak to podejście ma złożoność $O(n^2)$. Można zatem zmienić trochę podejście. A mianowicie:

1. Dane z każdego z z źródeł granulujemy na g granul liniowym algorytmem grupowania, np. FCM. Przyjmijmy, że z każdego źródła mamy n danych. To daje złożoność $O(zng)$.

2. Korzystamy z algorytmu k najbliższych sąsiadów, żeby znaleźć wartości brakujące. Złożoność: $O(z^2g^2)$.

3. Grupujemy uzupełnione wartości na l grup. Złożoność: $O(zgl)$.

4. Całkowita złożoność: $O(zng + z^2g^2 + zgl)$, co jest mniej niż $O(z^2n^2l)$, bo $g < n$.

Zadanie

Zadanie polega na empirycznym sprawdzeniu, czy zastosowanie takiego podejścia rzeczywiście skróci czas grupowania zachowując jednocześnie jakość grupowania.