

Marzena Nowakowska, Wydział Zarządzania i Modelowania Komputerowego, Politechnika Świętokrzyska

Wyodrębnianie grup jakości wody pod względem przydatności do spożycia z wykorzystaniem metody k-średnich

Zbiór danych *water_25-nr.csv* (nr – numer zespołu badawczego) zawiera zmienne będące miarami jakości wody. Z wyjątkiem cechy *Potability*, zmienne te mają charakter ilościowy. Zaznajomić się z opisem zmiennych w prezentacji wykładowej. Wyodrębnić w zbiorze skupienia, które charakteryzują stopień przydatności wody do spożycia. Zastosować algorytm podziału k-średnich dla podzbioru danych opisanego przez trzy wybrane przez siebie zmienne; do klasteryzacji pobrać znormalizować warianty tych zmiennych. Wykorzystać program *3_Skupienia_k-srednich.R*. W trakcie kolejnych etapów procesu badawczego należy zapoznać się i zrozumieć fragment kodu realizującego ten etap.

Raport zawiera nie więcej 4 strony.

Do każdego umieszczonego w raporcie wykresu należy się w treści odnieść.

Wybór zmiennych do analizy

Wybrać trzy zmienne do analizy skupień. Wyznaczyć dla nich miary statystyczne oraz krótko skomentować zawartość tabeli z tymi miarami.

Uwaga. Opis wykonać dla zmiennych oryginalnych.

Wybór liczby skupień

Zmodyfikować odpowiednią instrukcję programu, aby w analizie skupień były uwzględnione wybrane wcześniej zmienne. Wybrać trzy zmienne do analizy skupień. Wykorzystując wykresy osypiskowy dla WSS i wykres proporcji wyjaśnianej wariancji *BSS/TSS* określić docelową liczbę skupień.

Analiza skupień

Wykonać analizę skupień dla docelowej liczby klastrów (zmienna *k*). Ustalić wartość początkową do generatora; każdy zespół ma tę wartość inną – podać ją w raporcie. Sprawdzić znaczenie parametrów funkcji *kmeans*.

Zestawienie wyników

- Zestawić otrzymane wyniki zawarte w zmiennej programu *km*. Wykorzystać informacje zawarte w slajdzie 15 wykładu.
- Utworzyć wykres sylwetek (zespół wybiera układ poziomy albo pionowy wykresu). Doprecyzować elementy opisowe wykresu.
- Zilustrować skupienia w układzie 2D. skomentować stopień objaśniania zmienności zbioru przez dwie pierwsze główne składowe i ocenić jego jakość.

Charakterystyka skupień

Dla każdego skupienia wyznaczyć:

- miary statystyczne zmiennych uwzględnionych w analizie i zestawić je w tabeli,
- rozkłady zmiennej *Potability* i zilustrować je na wykresie słupkowym (jeden wykres dla wszystkich skupień).

Zinterpretować profile skupień, odnosząc się do dopuszczalnych wartości wybranych do analizy miar jakości wody. Uwzględnić rozkłady zmiennej *Potability*. Podać dla każdego skupienia etykietę.