Eksploracja danych

Piotr Lipiński

Lista zadań nr 4 – Grupowanie danych

Zadanie 0. *(2 punkty)*
Kilka popularnych algorytmów grupowania danych dostępnych jest w pakiecie SciKit do Pythona. Zapoznaj się z nimi wykonując skrypt umieszczony w materiałach do wykładu.
a) Jakie znaczenie ma parametr n_init w implementacji algorytmu KMeans? Powtórz obliczenia z różnymi wartościami tego parametru i przeanalizuj wyniki.
b) Jakie znaczenie ma parametr threshold w implementacji algorytmu Birch? Powtórz obliczenia z różnymi wartościami tego parametru i przeanalizuj wyniki.
c) Jakie znaczenie ma parametr eps w implementacji algorytmu DBScan? Powtórz obliczenia z różnymi wartościami tego parametru i przeanalizuj wyniki.
d) Rozszerz skrypt tak, aby na rysunkach z wynikami algorytmu DBScan widoczne były także punkty danych nie przypisane przez algorytm do żadnej grupy.
e) Wyniki grupowania danych IRIS są pokazywane na rysunkach dla dwóch pierwszych cech. Zrób rysunki dla pozostałych par cech.
f) Rozszerz skrypt tak, aby oceniał każde wykonane grupowanie danych wskaźnikami poprawności grupowania, takimi jak Silhouette Coefficient, Dunn Index oraz Davies-Bouldin Index.

Wskazówka: Silhouette Coefficient jest dostępny w pakiecie SciKit (http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient), definicje pozostałych wskaźników można znaleźć w literaturze i należy zaimplementować samemu.

Zadanie 1. *(2 punkty)*
Zapoznaj się dokładniej z implementacją algorytmu BIRCH w SciKit (http://scikit-learn.org/stable/modules/clustering.html#birch).
a) Wykonaj grupowanie z niskią wartością parametru threshold (na przykład 0.25). Algorytm utworzy wówczas dużo grup danych (zazwyczaj więcej niż potrzeba).
b) Zrób nowy rysunek wyników algorytmu BIRCH. Zaznacz na nim pozycje centrów grup przypisanych do korzenia utworzonego drzewa grupowania (korzeń drzewa zapisany jest w birch.root_, zaś centra w birch.root_.centroids_) oraz punkty danych w kolorach odpowiadającym utworzonym grupom. Jeśli utworzone drzewo grupowania ma więcej poziomów, to zrób jeszcze rysunek dla grupowania na poziomie niżej niż korzeń.
c) Jakie znaczenie ma parametr n_clusters? Czym różni się n_clusters =None od n_clusters = 3?

Zadanie 2. *(2 punkty)*
Normalizacja danych to przeskalowanie oryginalnych zarejestrowanych wartości cech na wartości z przedziału [0, 1] dokonywane za pomocą przekształcenia $x' = (x - x_{min}) / (x_{max} - x_{min})$, gdzie x' to przeskalowana wartość, x to wartość oryginalna, zaś x_{min} i x_{max} to odpowiednio minimalna i maksymalna zarejestrowana wartość cechy (normalizacji dokonuje się dla każdej cechy osobno).

Standaryzacja danych to przeskalowanie oryginalnych zarejestrowanych wartości cech na wartości o rozkładzie N(0, 1) dokonywane za pomocą przekształcenia $x' = (x - m) / s$, gdzie x' to przeskalowana wartość, x to wartość oryginalna, zaś m i s to odpowiednio średnia i odchylenie standardowe zarejestrowanych wartości cechy (standaryzacji dokonuje się dla każdej cechy osobno).

Czy normalizacja lub standaryzacja danych może mieć wpływ na działanie algorytmu K-Means? Jeśli tak, to skonstruuj proste przykłady pokazujące, że dane znormalizowane lub ustandaryzowane są lepiej grupowane niż dane oryginalne. Zrób stosowne rysunki. A jak jest w przypadku algorytmów BIRCH i DBScan?

Zadanie 3. *(4 punkty)*
W pliku APEX_OSD_V1_calibr_cube znajduje się wielospektralne zdjęcie satelitarne. Zdjęcie ma rozmiar 1500 x 1000 pikseli, każdy piksel jest opisany przez 285 wartości. Pogrupuj piksele tego zdjęcia używając poznanych algorytmów grupowania danych. Sprawdź różne algorytmy i różne ich parametry, uzyskane grupowania oceń stosując wybrane wskaźniki poprawności. Wyniki przedstaw w formie mapy terenu. Zastanów się, która z uzyskanych map jest najlepsza.

Wskazówki:

1. Plik APEX_OSD_V1_calibr_cube znajduje się w folderze /pio/scratch/2/ED2015/APEX/ APEX_OSD_Package_1.0/APEX_OSD_Package_1.0 dostępnym w systemie linux na (prawie wszystkich) komputerach w pracowniach 110 i 137.
2. W Matlabie wielospektralne zdjęcie satelitarne można wczytać poleceniem multibandread z odpowiednimi parametrami. W przypadku zdjęcia APEX można użyć poniższego kodu:

```
ImageRaw = multibandread('APEX_OSD_V1_calibr_cube', [1500, 1000, 285],
'int16', 0, 'bsq', 'ieee-le');
```

UWAGA: Obliczenia mogą być czasochłonne i wymagać dużej ilości pamięci. Sugeruje pracować na komputerach w pracowni 110.