K-Means, DBSCAN, Algorytm hierarchiczny

Otwórz w Wece plik iris.arff.

Iris Setosa

Iris Versicolor

Iris Virginica

- Przejdź na zakładkę **Visualize** i zapoznaj się z rozkładem danych:

  - Są to trójwymiarowe wykresy bazujące na trzech wybranych atrybutach ze zbioru danych: oś X, oś Y i różne kolory - upewnij się, że kolorem oznaczony jest atrybut class (domyślnie).
  - Kliknięcie na wykres pozwala na jego powiększenie (na powiększeniu można manipulować atrybutami położonymi na osiach).

Wyobraź sobie sytuację, w której znika atrybut `class` (znikają kolory) – czy jesteś w stanie ponownie pogrupować te dane w pierwotne klasy?

- Spróbuj to zrobić za pomocą algorytmu K-Means zaimplementowanego w Wece:

  1. Przejdź do zakładki **Cluster**.
  2. Zostaw domyślną wartość `Cluster mode` = `Use training set`
  3. Wybierz algorytm **SimpleKMeans**, w ustawieniach zmień liczbę klastrów na 3 i uruchom go.
  4. Zapoznaj się z wynikami:
     - Z wyświetlonym raportem
     - Z wizualizacją: kliknij prawym na wynik klasteryzacji w `Result list` i z menu wybierz `Visualize cluster assignments`
     - Przede wszystkim zapoznaj się z wizualizacją: X = Class, Y = Cluster. Możesz jeszcze pobawić się suwakiem `Jitter`, który rozrzuca dane (bo wiele punktów znajduje się w tym samym miejscu).
     - Czy dane zostały poprawnie zaklasyfikowane? Jaki błąd popełniliśmy?

- Spróbujmy to zrobić po raz kolejny:

  1. Kliknij przycisk `Ignore attributes`, zaznacz odpowiedni atrybut (który?) i kliknij `Select`. Uruchom ponownie klasteryzację.
  2. Zapoznaj się z wynikami i porównaj je z uzyskanymi wcześniej. Jaka teraz była skuteczność klasyfikacji?
  3. Czy wiesz który z trzech gatunków Irysów został w 100% poprawnie zaklasyfikowany przez ten algorytm? Możesz to odczytać z wykresu wizualizującego wyniki

klasyfikacji, jak również możesz w `Cluster mode` zaznaczyć opcję `Classes to clusters evaluation`. Po wybraniu tej opcji i wyświetleniu wizualizacji możesz zobaczyć na wykresie krzyżyki (poprawne trafienia) i kwadraty (niepoprawne klasyfikacje).

- Przeprowadź klasteryzację po raz trzeci. Tym razem nie zmieniaj żadnych ustawień. Czy pojawiły się dokładnie takie same wyniki?

- Jak to było wcześniej powiedziane, algorytm K-Means losuje początkowe położenie centroidów, więc wyniki powinny się od siebie różnić. Tutaj jednak są takie same…
- W jaki sposób sprawić aby wylosowały się inne wyniki?
- Przeprowadź klasteryzację za pomocą algorytmu DBSCAN i al. Hierarchicznego, dla tych samych danych.

DBSCAN: Density-based spatial clustering of applications with noise. Algorytm łączy w klastry obserwacje leżące blisko siebie (wg zasady sąsiad mojego sąsiada należy do tej samej grupy, co ja). Algorytm ignoruje pojedyncze punkty lub niewielkie skupiska.

Dwa paramtery:

•epsilon -jeśli dla dwóch punktów odległość $d(x1,x2)<epsilon$, to należą do tej samej grupy

•min_points–ignorowane są grupy $|Ci|<=$min$_points$(zawierające mniej niż min$_points$ obserwacji)

Algorytm hierarchiczny (aglomeracyjny),

Ważne elementy, decydujące o jego działaniu:

- Dobór metryki
- Metoda łączenia grup