AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 4

Algorytmy grupowania danych

data aktualizacji: 11.05.2022

Cel ćwiczenia

Celem ćwiczenia jest zapoznanie się z wybranymi algorytmami uczenia nienadzorowanego w ramach zadania grupowania (ang. *clustering*, czasem zwane klasteryzacją). Treść zadania obejmuje pracę z dwoma algorytmami grupowania k-means (k-średnich) oraz PAM (*Partition Around Medoids*). W ramach realizacji zdania zbadane zostaną różne konfiguracje algorytmów, użyte zostaną 3 miary oceny jakości grupowania oraz metody wizualizacji danych.

Celem pośrednim ćwiczenia jest zapoznanie się z systemem (platformą oraz językiem) R w zadaniu grupowania.

Wprowadzenie

Uczenie nadzorowane (lub uczenie z nauczycielem) związane jest z tworzeniem modelu na podstawie danych, który dokładnie wie jakie jest wejście i wyjście modelu (np. w klasyfikacji mamy oznaczenie klasy dla każdego rekordu), a algorytm w procesie uczenia budując model wykorzystuje tę wiedzę. W uczeniu nienadzorowanym brakuje takiej informacji (brak

nauczyciela…) i właśnie z taką sytuacją mamy do czynienia w zadaniu grupowania. Mamy dane (zbiór rekordów), dla których brak jest oznaczeń (klas), a zadanie sprowadza się do ich pogrupowania.

Przy zadaniu grupowania pojawiają się pytania. Jak pogrupować? Co znaczy dobre pogrupowanie danych? Czy powinno być dużo klastrów? Czy klastry powinny skupiać tylko podobne dane? Jak bardzo klastry powinny być „daleko" w przestrzeni danych? Na te pytania odpowiedzi są różne, w zależności od analizowanego zbioru danych oraz… przyjętej miary jakości klasteryzacji.

Co jeśli chcemy pogrupować dane, które mamy oznaczone (klasą)? Wtedy w zadania grupowania dodatkowo należy przeanalizować jak klasy rozłożyły się w klastrach. Jak klastry są jednorodne (miara *Purity*)? Ile (%) klas znajduje się w danym klastrze?

Użyte zbiory danych

Zbiór danych (łatwy): IRIS

Zbiory danych (trudniejsze): WINE, GLASS

Zbiór „wdzięczny" do grupowania:

SEEDS https://archive.ics.uci.edu/ml/datasets/seeds

Sugerowane narzędzia

Narzędzia sugerowane do realizacji tego zadania są ogólnodostępne, w większości na licencji *opensource*. Do realizacji zdania potrzebne będzie zainstalowanie środowiska R (zrealizowane w poprzednim zadaniu). Dodatkowe źródła edukacyjne odnośnie samego R:

http://books.goalkicker.com/RBook/

http://www.biecek.pl/R/ → start z R

https://www.rdocumentation.org/

Przykładowe biblioteki

Grupowanie:

k-means: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans

PAM: https://www.rdocumentation.org/packages/cluster/versions/2.1.0/topics/pam

Wizualizacja:

ggpubr

https://cran.r-project.org/web/packages/ggpubr/index.html

factoextra

https://cran.r-project.org/web/packages/factoextra/index.html

Miary grupowania:

Davies-Bouldin Index, Silhouette, Dunn https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Uwaga! Część miar jest minimalizowana, pozostałe są maksymalizowane. Warto zwrócić uwagę, jak poszczególne miary zachowują się w granicznych warunkach. Na przykład, co się dzieje gdy mamy tylko jeden klaster, lub co się dzieje jak mamy tyle samo klastrów co rekordów?

Miara Purity:

https://www.rdocumentation.org/packages/NMF/versions/0.17/topics/purity

https://www.rdocumentation.org/packages/funtimes/versions/7.0/topics/purity

Przebieg ćwiczenia

1. Sprawdzenie (+ew. instalacja) bibliotek związanych z algorytmami k-means, pam i wizualizacją danych.
2. Wczytanie danych i wstępna wizualizacja
3. Uruchomienie pam/k-means na prostym zbiorze (IRIS).
4. Wizualizacja danych, sprawdzenie zachowania miar klasteryzacji dla różnych konfiguracji
5. Sprawdzenie miary *Purity* dla różnych konfiguracji pam/k-means na prostym zbiorze (iris)

6. Przebadanie algorytmu PAM na 3 „trudniejszych” zbiorach. Zbadanie jak zmieniają się wartości miar przy zmianie parametrów:
    - k – liczba klastrów
    - stand – standaryzacja danych
    - matric – miara odległości danych
7. Wizualizacja wyników algorytmu PAM
8. Przebadanie algorytmu k-means na 3 „trudniejszych” zbiorach. Zbadanie jak zmieniają się wartości miar przy zmianie parametrów:
    - nstart – liczba n-startów
    - iter.max – maksymalna liczba iteracji
    - centers – początkowa liczba klastrów
9. Wizualizacja wyników algorytmu k-means
10. Sprawdzenie jak miara Purity opisuje wyniki 3 zbiorów znalezione w poprzednich etapach zadania.

Punktacja

Przy realizacji zadania student może otrzymać **max 10 punktów** wedle poniższej tabeli.

2	Zapoznanie się z bibliotekami R do klasteryzacji
2	K-means – zbadanie algorytmu, wstępne strojenie 3 wartości parametrów dla 3 zbiorów (glass, wine, seeds)
2	PAM – zbadanie algorytmu, wstępne strojenie 3 wartości parametrów dla 3 zbiorów (glass, wine, seeds)
2	Analiza jakości obu algorytmów w kontekście miary *Purity*
2	Porównanie działania algorytmów wraz z wizualizacją wyników (wykresy)

W zadaniu można narysować milion wykresów i milion tabelek. Założenie sprawozdania jest ukazanie tylko pewnego aspektu przeprowadzonych badań (nie wszystkie!). W tym celu podawane są powyższe punkty realizacji zadania oraz pytania pomocnicze.

Przy realizacji tego zadania wystarczy prosty raport PDF, z wykresami i ew. tabelkami. Przy wynikach badań należy podać komentarz, podać wnioski i podsumowanie.

Pytania pomocnicze

1. Czy przy grupowaniu potrzebna jest normalizacja/standaryzacja danych?
2. Co różni oba algorytmy z punktu widzenia reprezentacji klastra?
3. Który z algorytmów jest mniej odporny na szum i wartości odstające (ang. *outliers*)? Dlaczego?
4. Czy w zadaniu grupowania powinniśmy użyć walidacji krzyżowej?
5. Czy wyniki badanych algorytmów klasteryzacji powinny być powtarzane i uśredniane?
6. Co mierzą miary klasteryzacji podane w treści zdania?
7. Czy podejście aby liczba klastrów i liczba klas byłą taka sama jest poprawne w kontekście miary Purity?

Literatura

1. Wykłady do przedmiotu autorstwa prof. H.Kwaśnickiej
2. Cichosz P. ”Systemy uczące się”, WNT Warszawa
3. Zasoby Internetu: uczenie maszynowe (machine learning), data mining, R, grupowanie, klasteryzacja, clustering, k-means, pam