Eksploracja danych
Lista nr 4
27 V 2025

Zadanie 1 Zaawansowane metody klasyfikacji (kontynuacja zad.2 z listy nr 3)

Celem analizy jest zastosowanie (wybranych) zaawansowanych metod klasyfikacji. Zadanie to jest kontynuacją zad.2 z listy 3 i wykonujemy je dla uprzednio wybranych danych.

a) **Rodziny klasyfikatorów/uczenie zespołowe (ang. ensemble learning)**

- Zastosuj przynajmniej dwa algorytmy typu *ensemble learning* (tzn. takie metody jak bagging, boosting i random forest) do konstrukcji reguł klasyfikacyjnych oraz zbadaj ich dokładność[1]. Czy otrzymujemy istotną redukcję błędu klasyfikacji w porównaniu do klasyfikatora bazowego (tj. pojedynczego drzewa klasyfikacyjnego)? Czy występują istotne różnice w dokładności klasyfikacji pomiędzy różnymi metodami konstrukcji klasyfikatorów złożonych?

b) **Metoda wektorów nośnych (SVM)**

- Wykorzystując algorytm SVM, zbuduj klasyfikatory dla różnych funkcji jądrowych (np. jądro liniowe, wielomianowe i radialne) i porównaj ich skuteczność. Czy i w jakim stopniu wybór funkcji jądrowej oraz wybór parametru kosztu C wpływa na dokładność metody SVM?
- Dla jądra radialnego postaraj się „dostroić" jednocześnie parametry γ i C. Porównaj dokładność klasyfikacji dla modelu z optymalnie wybranymi parametrami oraz modelu skonstruowanego dla domyślnych parametrów. Czy optymalizacja parametrów pozwoliła na poprawę skuteczności skonstruowanego klasyfikatora?

c) **Porównanie skuteczności metod**

- Porównując wyniki uzyskane w punktach a) i b), postaraj się rozstrzygnąć, która z rozważanych metod wykazała najlepszą skuteczność.

Zadanie 2 Analiza skupień – algorytmy grupujące i hierarchiczne

Celem zadania jest zastosowanie poznanych algorytmów analizy skupień (ang. clustering) oraz ocena i porównanie jakości grupowania. Przeprowadzając analizę, należy zastosować przynajmniej jedną wybraną metodę grupującą (np. algorytm PAM) oraz przynajmniej jedną metodę hierarchiczną (np. algorytm AGNES).

a) **Wybór i przygotowanie danych**

- Do analizy wybieramy jeden ze zbiorów wymienionych w zad.2/lista 3. *Uwaga:* Można wybrać ten sam albo inny zbiór niż w przypadku porównywania metod klasyfikacji.
- Aby ułatwić wizualizację wyników, w przypadku „większych" danych losujemy podzbiór zawierający 200 rekordów (wierszy).
- Przed zastosowaniem metod analizy skupień usuwamy zmienną grupującą zawierającą etykietki klas (grup). Zmienna ta powinna być jednak później wykorzystana do oceny jakości grupowania ($\rightarrow$ zewnętrzne wskaźniki walidacyjne).
- Postaraj się rozstrzygnąć, czy konieczne będzie zastosowanie standaryzacji przed wyznaczeniem macierzy odległości/odmienności.

[1] Do oceny dokładności klasyfikacji można zastosować taki sam schemat, jak w przypadku zad.2 z listy nr 3

b) **Wizualizacja wyników grupowania**
Przyjmując liczbę skupień K równą rzeczywistej liczbie klas:

- Zilustruj otrzymane wyniki (podział na skupienia) na wykresach rozrzutu, zaznaczając (np. różnymi kolorami) przynależność do poszczególnych skupisk. Dodatkowo, uzupełnij wykresy o informacje nt. rzeczywistej przynależności obiektów do grup/klas (np. różne kolory mogą oznaczać różne skupienia, a różne symbole – przynależność do klas).
  - *Wskazówka 1:* Aby przedstawić dane na dwuwymiarowym wykresie rozproszenia, można wykorzystać odpowiednią metodę redukcji wymiaru (np. PCA lub MDS).
  - *Wskazówka 2:* W przypadku algorytmu hierarchicznego, aby otrzymać partycję dla ustalonej liczby skupień K można wykorzystać funkcję `cutree()`.
- W przypadku metody hierarchicznej porównaj dendrogramy otrzymane dla różnych metod łączenia skupień (ang. linkage methods) Dla chętnych: można zastosować także odpowiednie kolorowanie liści dendrogramów, zgodnie z rzeczywistą przynależnością obiektów do klas.
- Co można powiedzieć o podstawowych własnościach otrzymanych skupień (np. jednorodność/zwartość, separacja itp.)?
- Czy i w jakim stopniu otrzymany podział na klastry zgadza się z rzeczywistą przynależnością obiektów do klas?

c) **Ocena jakości grupowania. Wybór optymalnej liczby skupień i porównanie metod.**
Dla ustalonego zakresu liczby skupień ($K \in \{2, 3, \ldots, K.max\}$) wykonaj poniższe kroki:

- **wskaźniki wewnętrzne:** Wykorzystaj średnią wartość indeksu *silhouette* do porównania wyników otrzymanych dla różnych algorytmów analizy skupień (np. algorytmy PAM i AGNES) oraz różnej liczby skupień K.
- **wskaźniki zewnętrzne:** Wykorzystaj wybraną miarę zgodności dwóch partycji (np. funkcja `matchClasses(){e1071}`) do porównania wyników grupowania z rzeczywistą przynależnością do klas.

Na podstawie otrzymanych wyników spróbuj rozstrzygnąć, który algorytm lepiej poradził sobie z grupowaniem danych oraz jaka liczba klastrów jest optymalna.

d) **Interpretacja wyników grupowania – charakterystyki skupień**

- Wykorzystując wnioski z poprzedniego punktu, wyznacz podział na skupienia dla optymalnej liczby skupień K.
- Analizując podstawowe charakterystyki poszczególnych cech, zbadaj czym wyróżniają się obiekty należące do danego skupienia (np. dla obiektów należących do poszczególnych skupień można wyznaczyć średnie wartości cech, porównać wykresy pudełkowe itp.).
- W przypadku metody PAM sprawdź także, które obiekty są medoidami (reprezentantami skupień) i co je wyróżnia.