Eksploracja Danych

r.akad. 2014/15

Zadanie projektowe

Celem zadania jest przeprowadzenie analizy, grupowania i klasyfikacji na rzeczywistych danych z wykorzystaniem poznanych metod i narzędzi. Indywidualnie przypisany każdemu studentowi zbiór danych jest fragmentem pewnego standardowego zbioru danych (dostępnego w wersji oryginalnej pod adresem: https://archive.ics.uci.edu/ml/datasets/Glass+Identification). Zbiór danych SZKLO zawiera losowo wybrane próbki z większego zbioru wykorzystywanego oryginalnie przy prowadzeniu ekspertyz sądowych z zakresu kryminologii. Zbiór zawiera dane próbek szkła dotyczące ich parametrów fizykochemicznych:

1. Współczynnik załamiania światła
2. Zawartość sodu (Na)
3. Zawartość magnezu (Mg)
4. Zawartość aluminium (Al)
5. Zawartosć krzemu (Si)
6. Zawartość potasu (K)
7. Zawartość wapnia (Ca)
8. Zawartość baru (Ba)
9. Zawartość żelaza (Fe)

Wykonaj analizę tych danych oraz przeprowadź na nich grupowanie i klasyfikację. W szczególności wykonaj następujące zadania:

1. Określ liczbę obiektów, zakresy zmienności poszczególnych atrybutów, ich wartości średnie i odchylenia standardowe. Wskaż atrybuty o największej i najmniejszej zmienności zgodnie z miarami rozstępu i odchylenia standardowego. Jakie wnioski możesz wyciągnąć z tej analizy ?
2. Oceń czy, dla poszczególnych atrybutów, występują punkty oddalone.
3. Oceń wizualnie (analizując wykresy punktowe, macierz wykresów punktowych) na ile grup można podzielić zbiór danych wybierając wszystkie lub jedynie część spośród atrybutów. Odpowiedź uzasadnij.

4. Oceń czy wybrane atrybuty wymagają normalizacji lub standaryzacji. Jeśli tak, to wykonaj ją.
5. Dokonaj grupowania danych dla różnych liczb grup, znajdź – twoim zdaniem - optymalną liczbę grup. Czy w procesie grupowania konieczne jest wykorzystanie wszystkich atrybutów, czy wystarczy wybrać ich podzbiór ? Jeśli podzbiór to określ ten podzbiór i wykonaj dla niego grupowanie.
6. Utwórz wektor wartości atrybutu decyzyjnego na podstawie wyników grupowania
7. Podziel macierze atrybutów opisujących i atrybutu decyzyjnego na zbiór uczący i testowy.
8. Znajdź najlepszy klasyfikator spośród przerobionych w ramach ćwiczeń laboratoryjnych. Jako kryterium oceny przyjmij średni błąd klasyfikacji na zbiorze testowym.

Powyższe zadania wykonaj w środowisku R wykorzystując poznane w trakcie wykonywania poleceń wg. instrukcji, wbudowane oraz otrzymane od prowadzącego funkcje.
Opisując w raporcie efekty analizy danych, podaj wyniki, tok rozumowania i wnioski, zamieść wykresy pokazujące najważniejsze aspekty analizy. Do raportu dołącz skrypt(y) zawierający wszystkie komendy (wraz z komentarzami), które wykonałeś. Skrypt będzie wykorzystany do weryfikacji danych z raportu. Raport wraz ze skryptem należy wysłać pod adres iwanowski@ee.pw.edu.pl do końca semestru zimowego. W nagłówku wiadomości proszę wpisać „ED projekt Eksploracja, imię i nazwisko” (podając oczywiście swoje imię i nazwisko). W ciągu najpóźniej 24 godzin od otrzymania mejla będę ten fakt potwierdzał nadawcy. W razie nieotrzymania potwierdzenia proszę o informację.

Zadanie nie posiada jednego właściwego rozwiązania „wzorcowego”. Oceniany będzie opisany w sprawozdaniu tok rozumowania i otrzymane wyniki. Raport będzie weryfikowany poprzez uruchomienie skryptu. Dlatego proszę o wyczerpujący opis w sprawozdaniu oraz o czytelny, zawierający niezbędne komentarze kod skryptu. W razie pytań lub wątpliwości proszę o kontakt mejlowy.

Marcin Iwanowski
iwanowski@ee.pw.edu.pl