Eksploracja Danych

Projekt

Zespół 13
Damian Wojciuk
Paweł Duda

Cel projektu

Celem zadania jest przeprowadzenie analizy, grupowania i klasyfikacji na zbiorze danych z wykorzystaniem poznanych metod i narzędzi. W szczególności wykonaj następujące zadania:

1. Określ liczbę obiektów, liczbę klas, zakresy zmienności poszczególnych atrybutów, ich wartości średnie i odchylenia standardowe dla całego zbioru i w poszczególnych klasach. Wskaż atrybuty o największej i najmniejszej zmienności zgodnie z miarami rozstępu i odchylenia standardowego. Jakie wnioski możesz wyciągnąć z tej analizy ?

2. Oceń wizualnie (analizując wykresy punktowe, macierz wykresów punktowych) czy podział na grupy reprezentowany przez atrybut decyzyjny odpowiada naturalnym skupieniom danych w przestrzeni atrybutów.

3. Przetestuj dostępne klasyfikatory, oceń czy do poprawnej klasyfikacji należy wykorzystać wszystkie atrybuty, czy wystarczy ich podzbiór ? Oceń czy wybrane atrybuty wymagają normalizacji lub standaryzacji. Jeśli tak, to wykonaj ją.

4. Przyjmij sensowną miarę jakości klasyfikacji i znajdź zgodny z nią najlepszy klasyfikator.

5. Dokonaj grupowania danych pomijając atrybut decyzyjny. Wykonaj grupowanie dla różnych liczb grup, znajdź – twoim zdaniem - optymalną liczbę grup. Czy w procesie grupowania konieczne jest wykorzystanie wszystkich atrybutów, czy wystarczy wybrać ich podzbiór ? Czy otrzymany podział jest zgodny z podziałem na klasy zawartym w atrybucie decyzyjnym ? Czy jest zgodny z wnioskami otrzymanymi w punkcie 2

Zadanie 1

Początkowo wykonaliśmy podstawową analizę zbioru danych. Aby określić liczbę obiektów, oraz klas użyliśmy wbudowanej funkcji “str()”.

Wynik funkcji:

```
str(dane)
'data.frame':    498 obs. of  10 variables:
 $ V1   : num  0.49 2.52 12.25 4.88 2.77 ...
 $ V2   : num  25.61 15.37 -11.88 8.16 8.73 ...
 $ V3   : num  -1.8 -12.56 -19.16 5.7 -5.43 ...
 $ V4   : num  -16.5 -23.5 21.9 -30.7 -26.6 ...
 $ V5   : num  32 36.9 6.8 -28.1 38.7 ...
 $ V6   : num  -3.99 -5.41 -18.01 -19.1 -6.15 ...
 $ V7   : num  4.13 13.8 -23.25 13.73 5.2 ...
 $ V8   : num  30.6 43.85 7.86 28.12 48.42 ...
 $ V9   : num  46 33.3 27.8 33.9 28.6 ...
 $ klasa: Factor w/ 3 levels "A","B","C": 1 1 3 2 1 1 2 1 2 3 …
```

W zestawi 9 znajduje się 498 obiektów o 9 atrybutach, podzielonych na 3 klasy.

```
>eksploruj(dane)
[1] PODSUMOWANIE DANYCH
[1] Podstawowe miary
       V1                V2                V3                V4                V5                V6                V7
 Min.   :-21.899   Min.   :-42.568   Min.   :-26.01770   Min.   :-51.945   Min.   :-46.533   Min.   :-38.138   Min.   :-66.247
 1st Qu.: -2.750   1st Qu.:-12.210   1st Qu.:-11.19258   1st Qu.:-33.824   1st Qu.:-28.608   1st Qu.:-23.298   1st Qu.:-33.174
 Median :  1.745   Median :  8.011   Median : -5.66079   Median :-16.064   Median : -9.802   Median :-12.762   Median :  2.395
 Mean   :  1.911   Mean   :  4.370   Mean   : -5.73044   Mean   : -7.016   Mean   : -6.013   Mean   :-14.071   Mean   : -7.924
 3rd Qu.:  6.718   3rd Qu.: 19.558   3rd Qu.:  0.03432   3rd Qu.: 28.979   3rd Qu.: 23.739   3rd Qu.: -5.595   3rd Qu.:  9.254
 Max.   : 26.691   Max.   : 43.600   Max.   : 13.11891   Max.   : 56.440   Max.   : 42.171   Max.   :  7.206   Max.   : 34.452
       V8              V9          klasa
 Min.   : 3.35   Min.   : 1.248   A:140
 1st Qu.:26.75   1st Qu.:19.300   B:201
 Median :33.79   Median :26.060   C:157
 Mean   :33.32   Mean   :26.487
 3rd Qu.:40.53   3rd Qu.:33.349
 Max.   :56.44   Max.   :59.592
…
```

Odchylenie standardowe:

v1 =7.173995 -minimalne
v2 =18.909221
v3 =7.451687
v4 =30.794309 -maksymalne
v5 =25.422645
v6 =10.461000
v7 =23.318748
v8 =9.665736
v9 =10.308942

Miary rozstępu:
v1 = 48.59050
v2 = 86.16788
v3 = 39.13662 -minimalna
v4 = 108.38434 -maksymalna
v5 = 88.70373
v6 = 45.34365
v7 = 100.69858
v8 = 45.34365
v9 = 100.69858

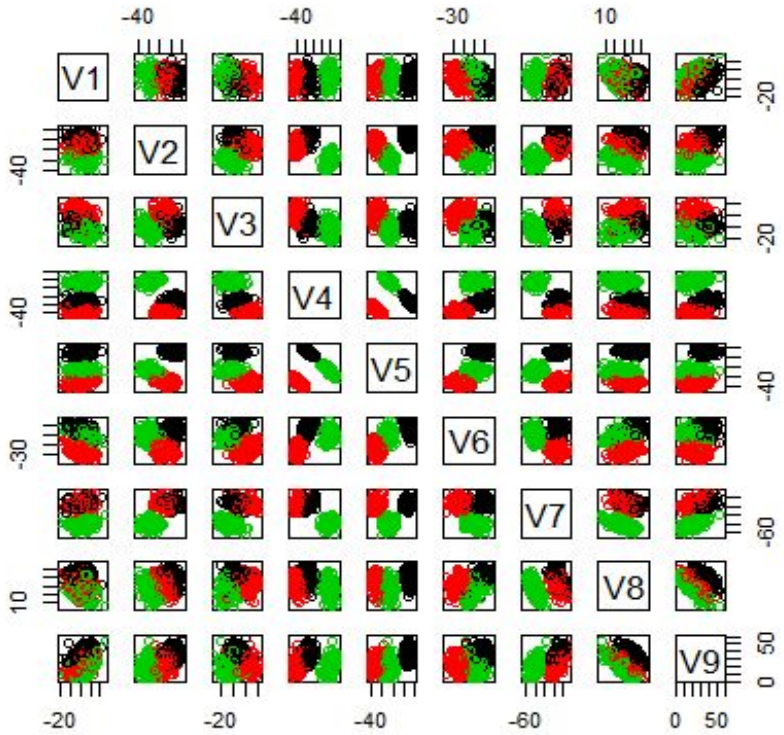
Wnioski:

Dane w V4 są najbardziej rozrzucone względem wartości średniej atrybutu, a w V3 najmniej oraz, że największy zakres zmienności posiada atrybut V4 a najmniejszy V1 .

Zadanie 2

Analizując macierz wykresów punktowych doszliśmy do następujących wniosków:

- podział na klasy wydaje się naturalny, a nie sztucznie wymuszony
- atrybuty V4 i V7 są silnie skorelowane



```
> cor(dane[,1:9])
          V1          V2           V3         V4         V5          V6          V7         V8          V9
V1  1.00000000 -0.22193884 -0.233264708  0.1643825 0.07704767 -0.14478821 -0.03406647 -0.2914412 0.326108984
V2 -0.22193884  1.00000000  0.335111771 -0.6598956 0.28212949  0.02289491  0.76849008  0.3106127 0.502878303
V3 -0.23326471  0.33511177  1.000000000 -0.7216079 -0.40359728 -0.43093699  0.53119359  0.1909120 0.008509498
V4  0.16438250 -0.65989556 -0.721607894  1.0000000 0.22003947  0.52222626 -0.84968779 -0.3477132 -0.143506006
V5  0.07704767  0.28212949 -0.403597278  0.2200395 1.00000000  0.79711188  0.05894627  0.3106629 0.505114177
V6 -0.14478821  0.02289491 -0.430936986  0.5222263 0.79711188  1.00000000 -0.26278872  0.3248442 0.238321872
V7 -0.03406647  0.76849008  0.531193592 -0.8496878 0.05894627 -0.26278872  1.00000000  0.2797333 0.495056928
V8 -0.29144115  0.31061268  0.190911969 -0.3477132 0.31066290  0.32484421  0.27973333  1.0000000 -0.212841304
V9  0.32610898  0.50287830  0.008509498 -0.1435060 0.50511418  0.23832187  0.49505693 -0.2128413 1.000000000
```

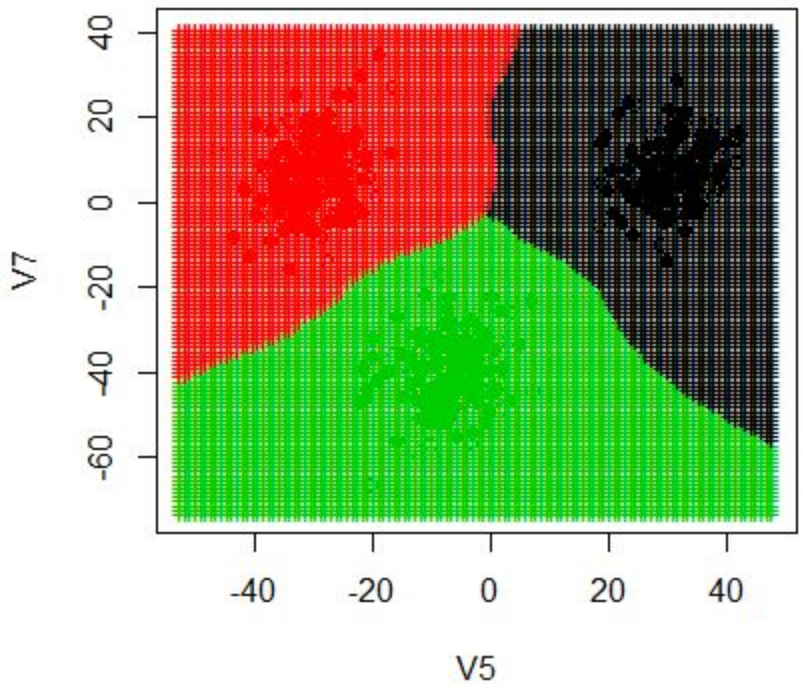
Wnioski:

Macierz wykresów punktowych pozwala w dogodny sposób analizować cały zbiór danych i tym samym umożliwia wyciągnięcie wniosków dotyczących słuszności dobranego podziału na grupy.

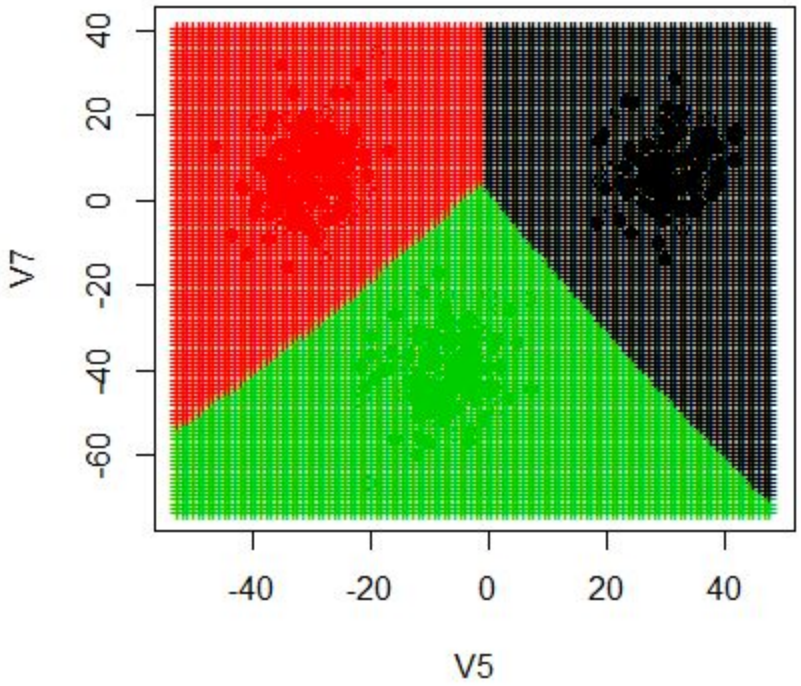
Zadanie 3

Na podst obserwacji wykresów z macierzy wykresów punktowych zbioru danych stwierdziliśmy, że atrybuty V5 i V7 będą najlepsze do klasyfikacji obiektów. Podział jest naturalny (klasy są wyizolowane) – potrzebujemy tylko tych 2 atrybutów

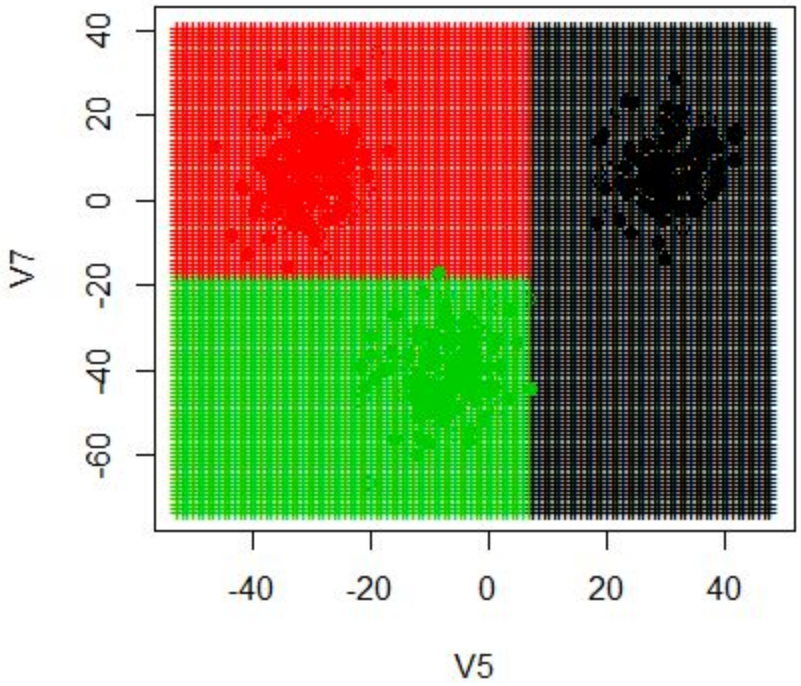
> granice("knn",dane,zu,5,7)



```
> granice("bayes",dane,zu,5,7)
```



```
> granice("drzewo",dane,zu,5,7)
```



Wnioski

Atrybuty nie wymagają normalizacji ani standaryzacji, ponieważ charakteryzują się zmiennością na podobnym poziomie.
Po przetestowaniu wszystkich klasyfikatorów i otrzymaniu za każdym razem podobnych wyników stwierdziliśmy, że knn będzie najlepszym wyborem. Jest najoszczędniejszy jeśli chodzi o zasoby pamięciowe i czasowe, a podczas klasyfikacji nie korzysta ze zbioru uczącego. Ponadto jego korzystną cechą w naszym przypadku jest brak fazy uczenia – w procesie klasyfikacji jest przeglądany cały zbiór uczący w poszukiwaniu najbliższego obiektu.

Zadanie4

“Klasyfikator jest dobry, jeśli 90% danych podlegających klasyfikacji będzie zgodna z klasyfikacją w zbiorze testowym.”

```
>weryfikuj("knn",dane,zu,c(5,7))

[1] Weryfikacja klasfikatora
[1]   zbiór uczący

    A   B   C
A  98   0   0
B   0 140   0
C   0   0 109
[1]   zbiór testowy

   A  B  C
A 42  0  0
B  0 61  0
C  0  0 48
```

```
>weryfikuj("bayes",dane,zu,c(5,7))

[1] Weryfikacja klasfikatora
[1]   zbiór uczący

     A   B   C
  A  98   0   0
  B   0 140   0
  C   0   0 109
[1]   zbiór testowy

    A  B  C
  A 42  0  0
  B  0 61  0
  C  0  0 48

>weryfikuj("drzewo",dane,zu,c(5,7))

[1] Weryfikacja klasfikatora
[1]   zbiór uczący

     A   B   C
  A  98   0   0
  B   0 140   0
  C   0   0 109
[1]   zbiór testowy

    A  B  C
  A 42  0  0
  B  0 61  0
  C  1  0 47
```

W przypadku drzewa decyzyjnego jeszcze lepszą dokładność uzyskamy w wyniku klasyfikacji dla atrybutów V4 i V5

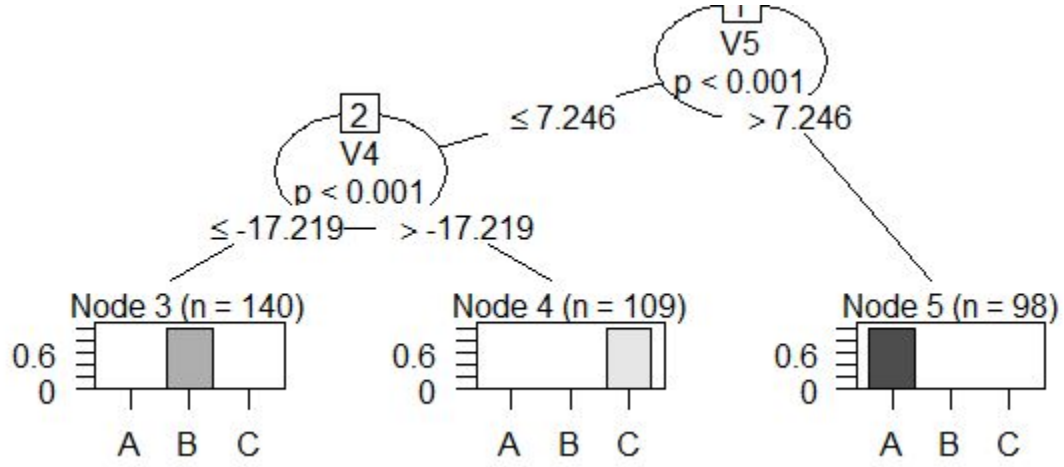
>weryfikuj("drzewo",dane,zu,c(4,5))

[1] Weryfikacja klasyfikatora
[1] zbiór uczący

	A	B	C
A	98	0	0
B	0	140	0
C	0	0	109

[1] zbiór testowy

	A	B	C
A	42	0	0
B	0	61	0
C	**0**	0	48

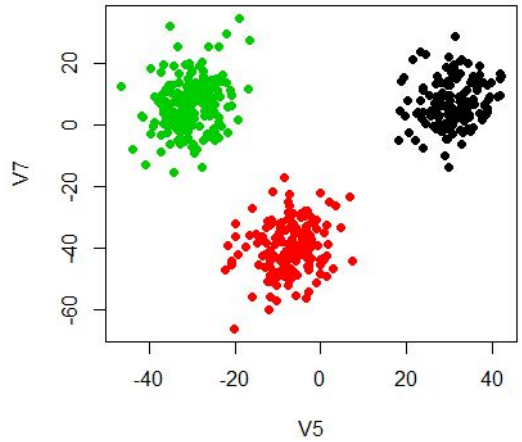


Wnioski:

Przed przystąpieniem do analizy przyjęliśmy miarę akceptacji klasyfikatora na poziomie 90% pokrycia. W wyniku klasyfikacji dla atrybutów V5 i V7 otrzymaliśmy ponad 99% pokrycia klas ze zbioru testowego z całym zbiorem danych, a w przypadku atrybutów V4 i V5 nawet 100%. Najlepszym klasyfikatorem w naszym przypadku okazał się knn i beyers.

‘

Zadanie 5



Wnioski:

Na podstawie obserwacji wynika, że optymalna jest początkowa liczba grup 3, przy innej ich liczbie nie można jednoznacznie określić tego podziału.

W procesie grupowania nie są potrzebne wszystkie atrybuty, wystarczy podzbiór (np. 2). Otrzymany podział pokrywa się z podziałem na klasy , ale na podstawie grupowania nie jesteśmy w stanie stwierdzić, która to konkretnie klasa. Jesteśmy jedynie w stanie określić, które obiekty należą do tej samej klasy). Otrzymane wnioski zgadzają się z tymi, do których doszliśmy w punkcie 2.