Eksploracja danych
Raport poprojektowy

Autorzy:
Adam Gryczka
Konrad Wyłucki

Spis treści

1 Wprowadzenie

Celem zadania jest przeprowadzenie analizy, grupowania i klasyfikacji na zbiorze danych z wykorzystaniem poznanych metod i narzędzi pakietu R.

2 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?

Liczbę obiektów w zbiorze danych określamy następująco:

```
> g <- read.table("zestaw_15.dat")
> length(g$V1)
[1] 697
```

Liczbę klas można określić w następujący sposób:

```
> length(unique(g$klasa))
[1] 3
```

2.1 Obliczanie parametrów bez podziału na klasy.

Zakres zmienności oraz wartość średnią parametrów bez podziału na klasy najłatwiej znaleźć w następujący sposób:

```
> summary(g)
       V1                V2                V3                V4
 Min.   :-25.970   Min.   :-4.253   Min.   :-12.35   Min.   :-45.381
 1st Qu.:-11.600   1st Qu.: 3.859   1st Qu.: 11.84   1st Qu.:-23.938
 Median : -4.526   Median : 7.065   Median : 23.71   Median : -5.276
 Mean   : -4.814   Mean   : 8.956   Mean   : 24.86   Mean   : -9.623
 3rd Qu.:  1.922   3rd Qu.:11.169   3rd Qu.: 38.47   3rd Qu.:  3.252
 Max.   : 19.762   Max.   :29.998   Max.   : 71.49   Max.   : 22.290

       V5                V6                V7                V8
 Min.   :-10.1950  Min.   :-33.694  Min.   :-32.472  Min.   :-41.775
 1st Qu.: -0.8769  1st Qu.: -5.291  1st Qu.: -5.076  1st Qu.:-16.445
 Median :  4.6048  Median :  3.588  Median :  4.915  Median : -6.848
 Mean   : 18.0935  Mean   :  2.895  Mean   :  6.142  Mean   : -7.204
 3rd Qu.: 44.2248  3rd Qu.: 11.345  3rd Qu.: 23.378  3rd Qu.:  2.494
 Max.   : 54.8004  Max.   : 38.890  Max.   : 40.522  Max.   : 19.492
```

```
       V9          klasa
 Min.   :-14.830   A:276
 1st Qu.:  1.295   B:153
 Median : 13.953   C:268
 Mean   : 11.601
 3rd Qu.: 21.170
 Max.   : 37.807
```

Odchylenie standardowe bez podziału na klasy znajdujemy w następujący sposób:

```
> sd(g$V1)
[1] 8.935198
> sd(g$V2)
[1] 7.223701
> sd(g$V3)
[1] 16.22782
> sd(g$V4)
[1] 15.73127
> sd(g$V5)
[1] 22.37508
> sd(g$V6)
[1] 11.82935
> sd(g$V7)
[1] 18.27164
> sd(g$V8)
[1] 12.00164
> sd(g$V9)
[1] 11.36592
```

2.2 Obliczanie parametrów z podziałem na klasy.

Żeby poprawić przejrzystość otrzymanych wyników zostały one przedstawione w tabelach.

V1:	**Min.**	**Max.**	**Wartość średnia**	**Odchylenie standardowe**
klasa A	-25.970	2.331	-12.633	5.702478
klasa B	-16.4709	9.1694	-4.7662	5.728267
klasa C	-9.8010	19.7624	3.2118	5.300615

V2:	**Min.**	**Max.**	**Wartość średnia**	**Odchylenie standardowe**
klasa A	-2.553	15.362	6.768	3.302345
klasa B	12.45	30.00	20.85	3.252166
klasa C	-4.253	17.660	4.418	3.469798

V3:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	13.93	71.49	40.84	8.576347
klasa B	1.099	45.909	21.921	9.405737
klasa C	-12.346	35.022	10.077	8.493822

V4:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	-45.3807	0.7894	-26.7122	7.268271
klasa B	-19.162	18.456	-1.638	7.621561
klasa C	-15.228	22.290	3.417	6.964821

V5:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	36.54	54.80	45.22	3.414899
klasa B	-5.712	14.610	4.209	3.609429
klasa C	-10.1950	7.0830	-1.9206	3.489449

V6:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	-14.403	38.890	10.912	8.639283
klasa B	-33.694	9.254	-10.366	8.917345
klasa C	-17.461	24.159	2.210	8.657917

V7:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	12.96	40.52	25.21	4.629303
klasa B	-32.472	-6.871	-20.757	5.237181
klasa C	-10.009	15.406	1.857	4.852666

V8:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	-41.775	-3.835	-18.682	6.792099
klasa B	-28.2574	9.7407	-5.3618	7.094758
klasa C	-16.785	19.492	3.565	6.690186

V9:	Min.	Max.	Wartość średnia	Odchylenie standardowe
klasa A	1.927	37.807	21.193	5.640984
klasa B	-2.26	33.25	15.60	6.086903
klasa C	-14.8299	14.9445	-0.5632	5.483617

2.3 Atrybuty o największej i najmniejszej zmienności.

Zmienność atrybutów bez podziału na klasy przedstawia się następująco:

- największy rozstęp - atrybut V3: 83.83368

- najmniejszy rozstęp - atrybut V2: 34.25131
- największe odchylenie standardowe - atrybut V5: 22.37508
- najmniejsze odchylenie standardowe - atrybut V2: 7.223701

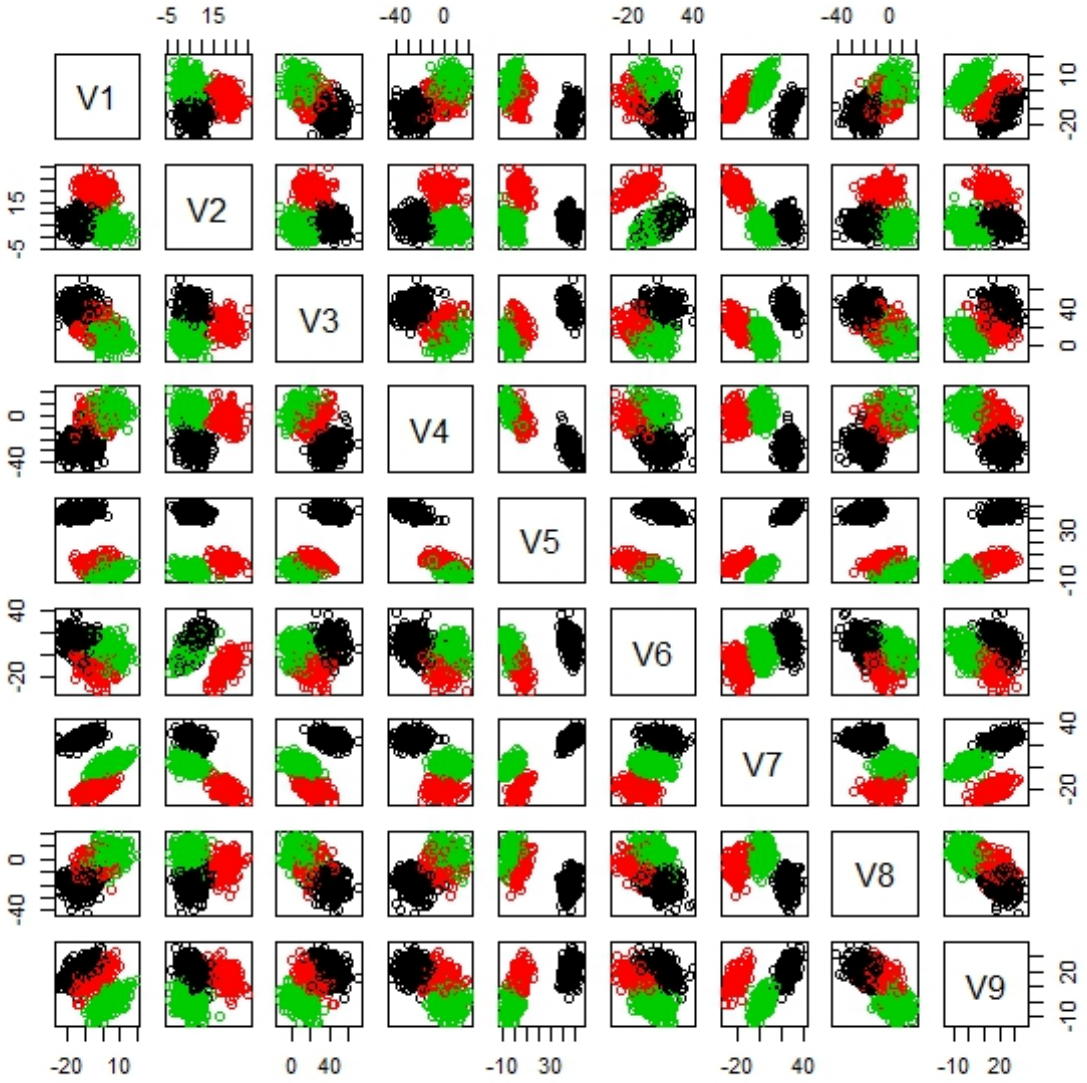
Zmienność atrybutów z podziałem na klasy:

1. Klasa A
    - największy rozstęp - atrybut V3: 57.55569
    - najmniejszy rozstęp - atrybut V2: 17.91452
    - największe odchylenie standardowe - atrybut V6: 8.639283
    - najmniejsze odchylenie standardowe - atrybut V2: 3.302345
2. Klasa B
    - największy rozstęp - atrybut V3: 44.80978
    - najmniejszy rozstęp - atrybut V2: 17.54532
    - największe odchylenie standardowe - atrybut V3: 9.405737
    - najmniejsze odchylenie standardowe - atrybut V2: 3.252166
3. Klasa C
    - największy rozstęp - atrybut V3: 47.36809
    - najmniejszy rozstęp - atrybut V5: 17.27802
    - największe odchylenie standardowe - atrybut V6: 8.657917
    - najmniejsze odchylenie standardowe - atrybut V2: 3.469798

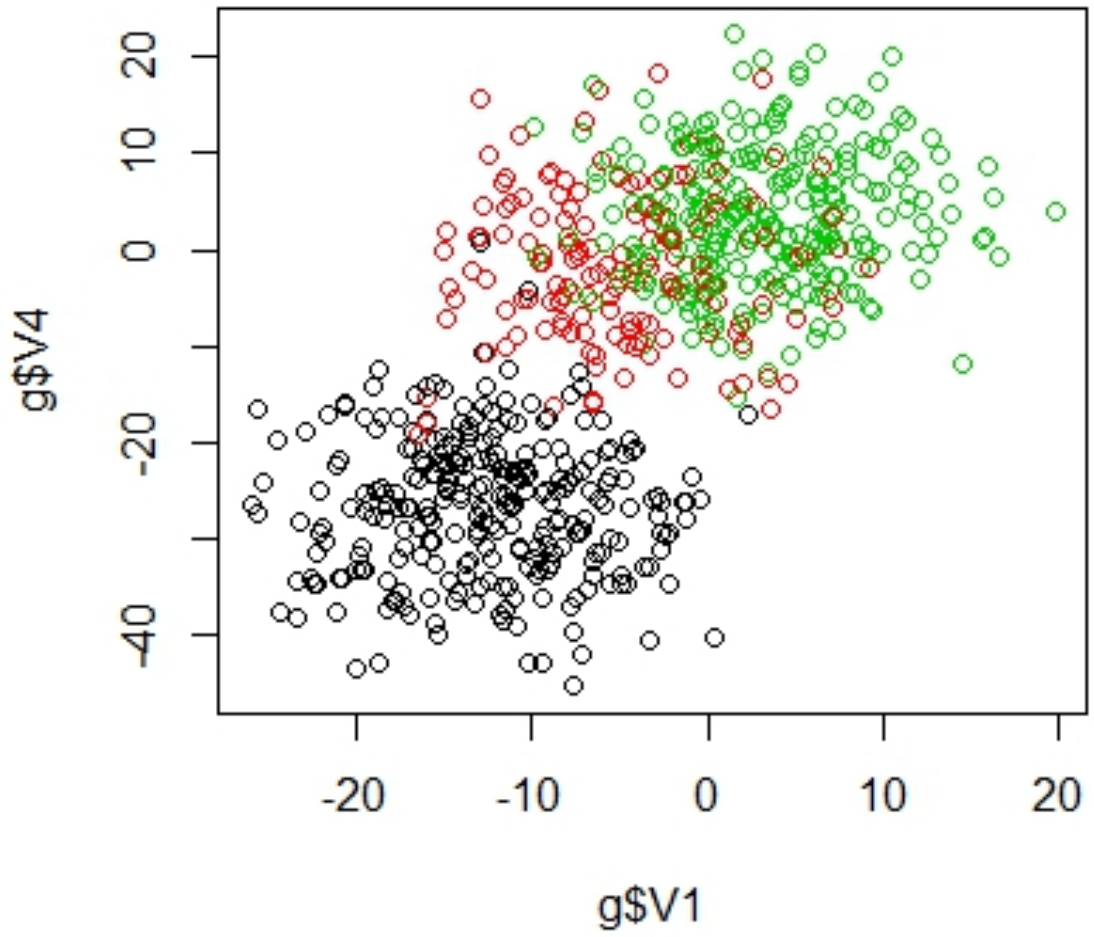
2.4 Wnioski

Wyróżniającym się atrybutem jest atrybut V2, którego cechy są charakterystyczne w zasadzie we wszystkich klasach. Ma on w większości klas najmniejsze rozstęp i odchylenie standardowe. Zauważyć należy, że wartości wszystkich atrybutów nie różnią się od siebie znacząco. Wszystkie watości mieszczą się w przedziale od -41,775 do 71,49.

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



Naszym zdaniem podział na grupy reprezentowany przez atrybut decyzyjny odpowiada naturalnym skupieniom danych w przestrzeni atrybutów. Zależnośći pomiędzy atrybutami w wielu przypadkach wyraźnie wskazują na granice oddzielające poszczególne grupy. Na niewielu wykresach grupy przenikają się w sposób uniemożliwiający odróżnienie ich od siebie. W takich przypadkach, najczęściej jedna z grup, dzieli atrybuty z jedną z pozostałych grup. Dobrze jest to widoczne na wykresie V1:V4.



4 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.

Badanie klasyfikatorów przeprowadziliśmy z użyciem dostarczonych funkcji `weryfikuj(...)` oraz `granice(...)`. Na początku wczytaliśmy dane z zewnętrznego pliku o nazwie `zestaw_15.dat`:

```
> g <- read.table("zestaw_15.dat")
```

Następnie wyodrębniliśmy w zbiorze `g` zbiory uczący oraz testowy:

```
> ug <- podziel(g)
```

W ten sposób mamy przygotowany zestaw danych do pracy. Następnie określamy zakresy zmienności poszczególnych atrybutów z użyciem dostarczonych funkcji:

```
> najwieksze(z, c(1:9)) - najmniejsze(z, c(1:9))
[1] 45.73281 34.25131 83.83368 67.67111 64.99538 72.58413 72.99433 61.26745
52.63676
```

Jak widać powyżej, każdy z przedziałów posiada dość duże różnice w przedziałach zmienności. Jedynie atrybuty 6, 7 posiadają zbliżone wartości przefdziałów zmienności. Jeżeli zatem dla nich przetestujemy klasyfikator minimalnoodległościowy (k=1), to otrzymamy następujący wynik:

```
> weryfikuj('knn', g, ug, c(6,7))
[1] Weryfikacja klasfikatora
[1]   zbiór uczący

     A   B   C
  A 193   0   0
  B   0 107   0
  C   0   0 187
[1]   zbiór testowy

    A  B  C
  A 83  0  0
  B  0 46  0
  C  0  0 81
```

Jest to bardzo dobry wynik. Wszystkie obiekty zostały prawidłowo sklasyfikowane. Skuteczność klasyfikatora wyniosła 100%. Ponownie wylosowaliśmy podzbiory uczący oraz testowy ze zbioru g jeszcze trzy razy, i skuteczność dla tych dwóch atrybutów wyniosła za pierwszym i drugim razem po 99.856528%, za trzecim razem ponownie 100%, a błędne klasyfikacje pojawiały się jedynie w macierzy kontyngencji zbioru testowego.
Z tego wniosek, że aby uzyskać skuteczność klasyfikacji na poziomie ponad 99%, wystarczy z naszego konkretnego zbioru danych pobrać dwa atrybuty (6 i 7) bez konieczności ich normalizacji, bądź standaryzacji.
Co jest zaskakujące, zdefiniowanie większej liczby sąsiadów daje wynik zazwyczaj równie satysfakcjonujący lub minimalnie gorszy (na poziomie 99.713056%):

```
> weryfikuj('knn', g, ug, c(6,7), k=4)
[1] Weryfikacja klasfikatora
[1]   zbiór uczący

     A   B   C
  A 193   0   0
  B   0 107   0
  C   1   0 186
[1]   zbiór testowy

    A  B  C
  A 83  0  0
  B  0 46  0
  C  1  0 80
```

Inne Klasyfikatory dają wynik porównywalny, i tak na przykład naiwny klasyfikator Bayesa (skuteczność: 99.5695839%):

```
> ug <- podziel(g)
> weryfikuj('bayes', g, ug, c(6,7))
[1] Weryfikacja klasfikatora
[1]   zbiór uczący

     A   B   C
  A 193   0   0
  B   0 107   0
  C   2   0 185
[1]   zbiór testowy

    A  B  C
  A 83  0  0
  B  0 46  0
  C  0  1 80
```

Następnie przetestowaliśmy różne klasyfikatory dla kilku różnych wektorów atrybutów opisujących (liczby w komórkach oznaczają błędnie sklasyfikowane obiekty odczytane z macierzy kontyngencji):

Nazwa klasyfikatora:	**knn (k=5)**	**np**	**bayes**	**drzewo (k=2)**
c(1,2)	43	51	44	46
c(3,4)	53	84	65	54
c(6,7)	0	2	4	4
c(1,2,4,5)	0	4	1	3
c(1,3,5,8,9)	0	4	4	6

Zgodnie z sugestią w poleceniu, zdecydowaliśmy się na znormalizowanie wszsytkich atrybutów:

```
> n <- normalizuj(g, 'norm', c(1:9))
```

i podział na zbiór uczący i testowy:

```
> un <- podziel(n)
```

Obliczamy (sprawdzamy), czy normalizacja przebiegła poprawnie:

```
> najwieksze(n, c(1:9)) - najmniejsze(n, c(1:9))
[1] 1 1 1 1 1 1 1 1 1
```

Po udanej normalizacji zbioru danych przeprowadzamy analogiczne testy klasyfikatorów, używając funkcji `weryfikuj(...)`:

```
> weryfikuj('knn', n, un, c(6,7))
[1] Weryfikacja klasfikatora
[1]   zbiór uczący

      A   B   C
  A 193   0   0
  B   0 107   0
  C   0   0 187
[1]   zbiór testowy

     A  B  C
  A 83  0  0
  B  0 45  1
  C  1  0 80
```

Inny przykład:

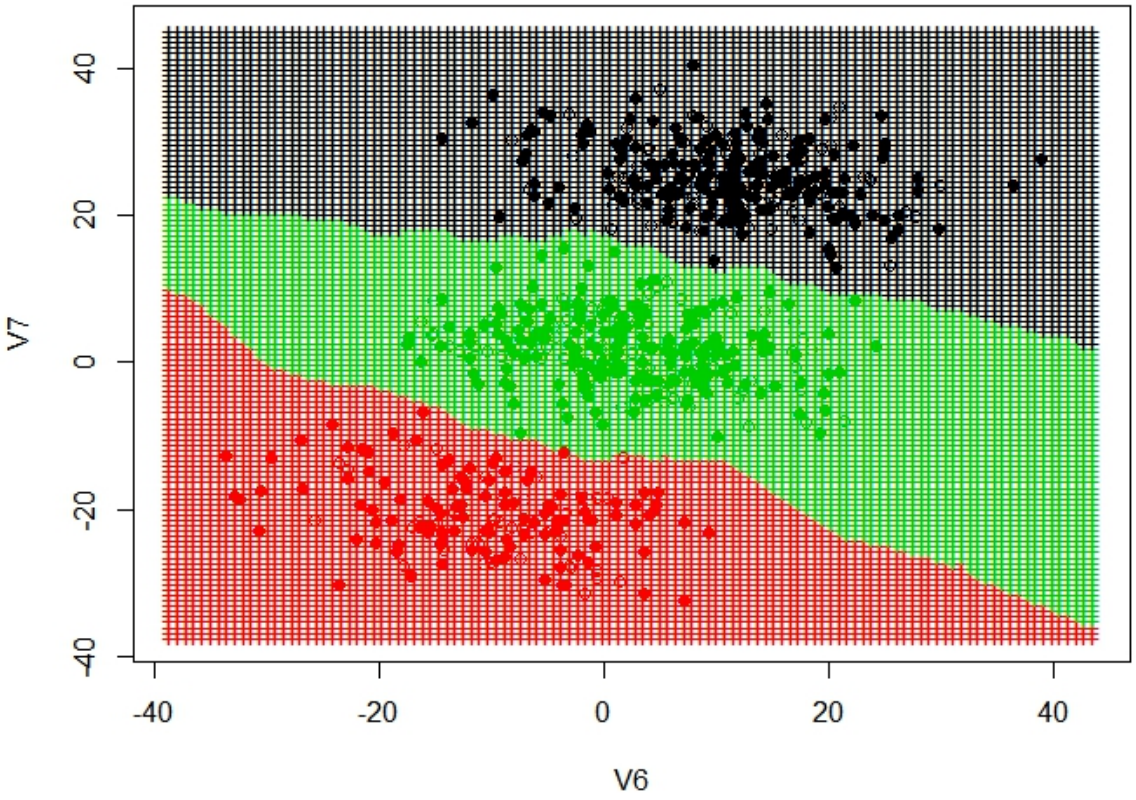
```
> weryfikuj('knn', n, un, c(6,7), k=4)
[1] Weryfikacja klasfikatora
[1]   zbiór uczący

      A   B   C
  A 193   0   0
  B   0 107   0
  C   0   1 186
[1]   zbiór testowy

     A  B  C
  A 83  0  0
  B  0 46  0
  C  1  0 80
```

Pozostałe porównania dają wynik podobny. Wnioskujemy z tego, iż normalizacja wejściowego zbioru danych przyniosła jedynie porównywalny (żeby nie rzec - gorszy) wynik, niźli analiza zbioru bez jego znormalizowania.

Ustandaryzowanie zbioru wejściowego również nie przynosi wymiernych korzyści. Ze względu na zacieranie czytelności tego dokumentu postanowiliśmy pominąć dowody ku temu i w celu ich przeanalizowania odsyłamy do wyników działania skryptu przeprowadzającego standaryzację i weryfikującego po tym działanie klasyfikatorów.

Pokażemy teraz przykładowy wykres, na którym widać zaznaczone obiekty, które przydzielono do odpowiednich klas:

```
> granice('knn', g, ug, 6, 7, k=5)
```



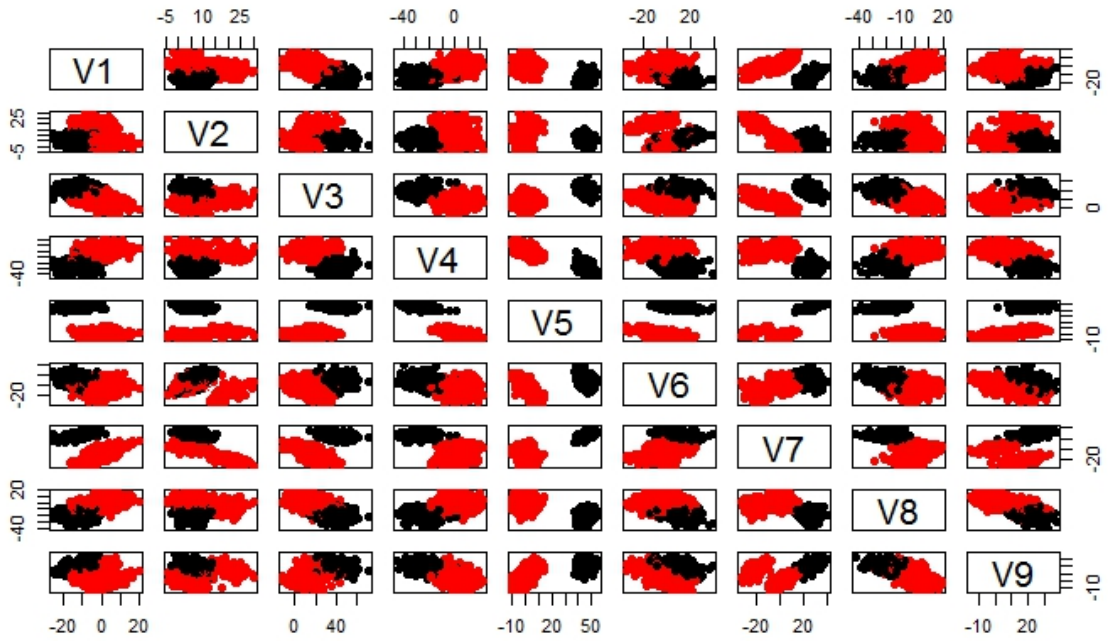
Rysunek 1: Zgodnie z powyższą tabelą, wszystkie obiekty zostały sklasyfikowane prawidłowo. Do klasyfikacji został użyty klasyfikator k-najbliższych sąsiadów, gdzie k = 5.

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

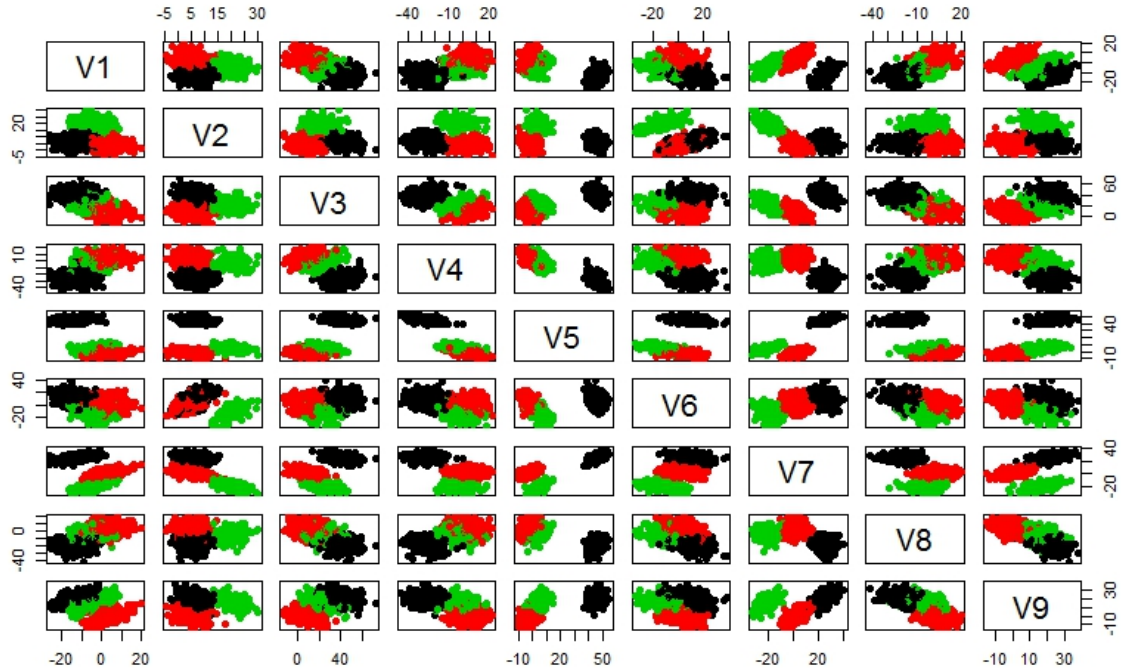
Według nas najlepszym sposobem mierzenia jakości klasyfikacji będzie macierz kontyngencji. Prezentuje ona w czytelny sposób to, na ile dany klasyfikator jest skuteczny. W odróżnieniu od wykresu, ma ona tę przewagę, że można łatwo obliczyć na przykład skuteczność klasyfikatora, podczas gdy obserwacja i zliczanie punktów na wykresie jest niewygodne dla małych zbiorów danych. Zgodnie z wyżej przeprowadzonymi testami klasyfikatorów najlepszym z nich okazał się klasyfikator k-najbliższych sąsiadów.

6 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 ?

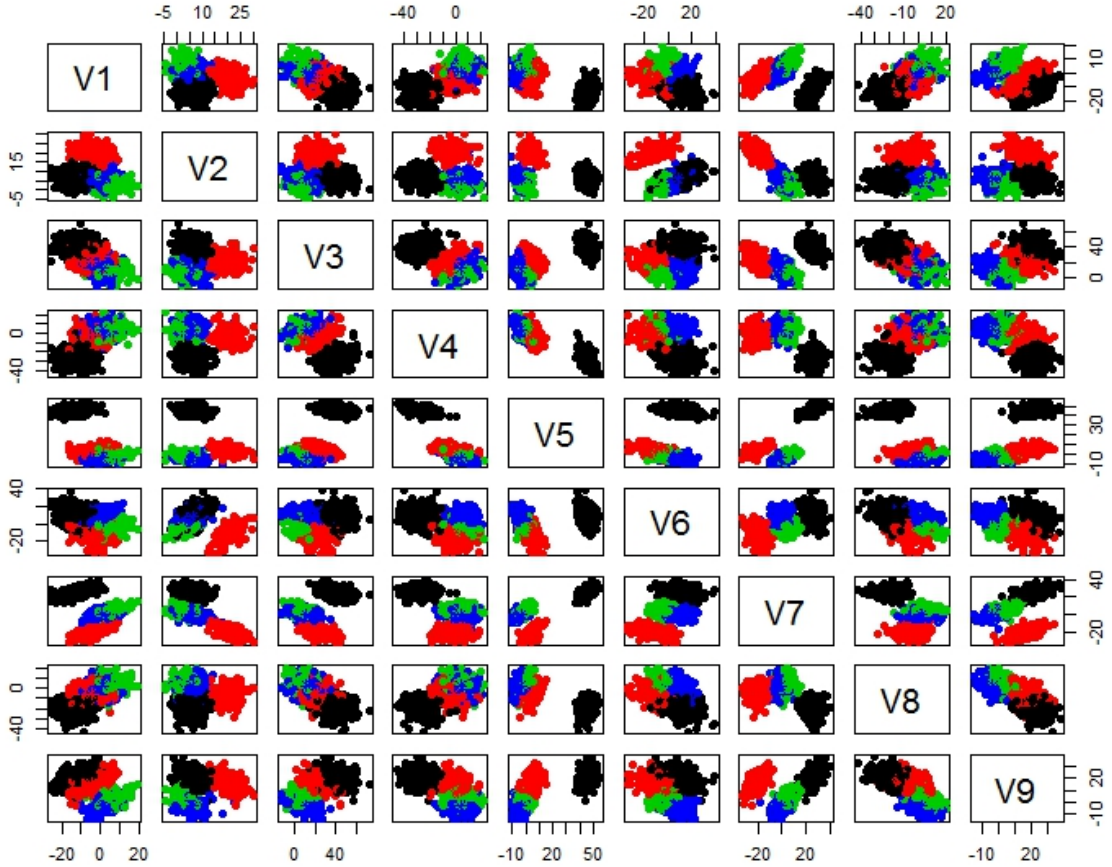
6.1 2 grupy.



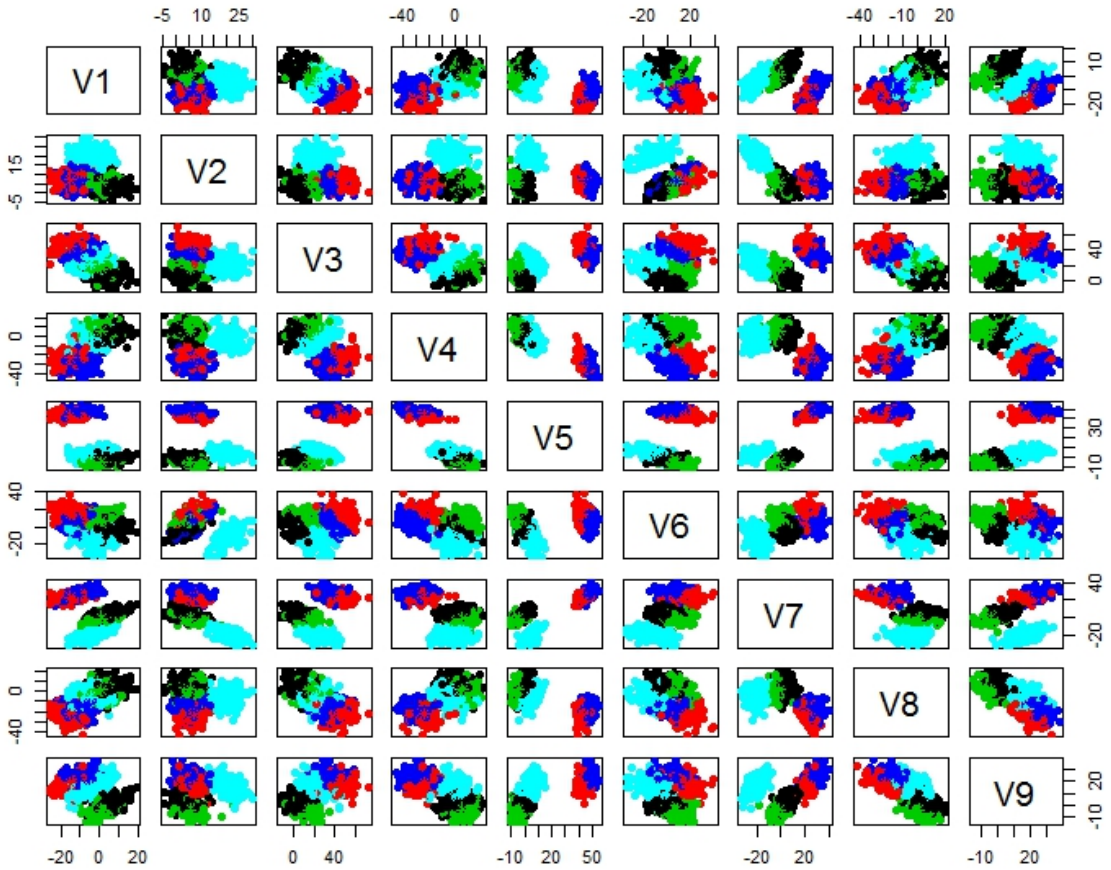
6.2 3 grupy.



6.3 4 grupy.



6.4 5 grup.



6.5 Wnioski

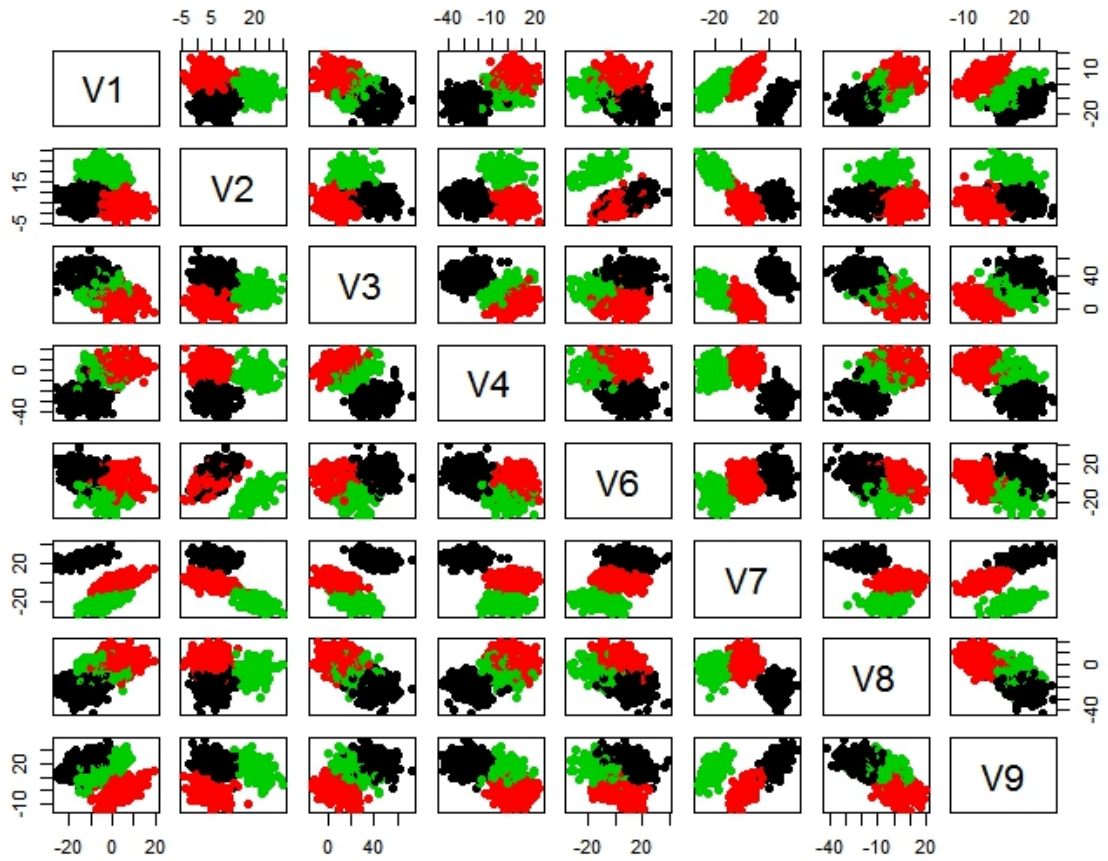
Odpowiednią liczbą grup wydaje się być liczba klas zaproponowana w zestawie danych, czyli 3. Podział na 2 grupy również wygląda dobrze, jednak w niektórych przypadkach obiekty należące do jednej grupy są zbyt odległe od siebie, co czyni podział nieoptymalnym. W przypadku większej liczby grup obiekty należące do różnych grup nachodzą na siebie, co czyni obie te grupy praktycznie niemożliwymi do rozróżnienia.

W procesie grupowania nie trzeba wykorzystywać wszystkich atrybutów. Dobrym rozwiązaniem jest pominięcie jednego z dwóch atrybutów, między którymi zachodzi zjawisko korelacji.

Oto macierz korelacji między wszystkimi atrybutami.

```
       V1           V2          V3          V4         V5
V1  1.0000000 -0.148614444 -0.68312835  0.67217403 -0.6985897
V2 -0.1486144  1.000000000 -0.00866217  0.08797492 -0.1728212
V3 -0.6831284 -0.008662170  1.00000000 -0.64724926  0.7830394
V4  0.6721740  0.087974924 -0.64724926  1.00000000 -0.9179388
V5 -0.6985897 -0.172821233  0.78303943 -0.91793884  1.0000000
V6 -0.3654767 -0.273815691  0.30926128 -0.46080307  0.4512856
V7 -0.3379662 -0.654987694  0.49034379 -0.69424261  0.8097183
V8  0.6598677  0.002022226 -0.72157453  0.66458129 -0.7589533
V9 -0.4618792  0.201709925  0.58990127 -0.69238338  0.7489092
       V6             V7          V8           V9
V1 -0.3654767217 -0.3379662  0.659867731 -0.4618791829
V2 -0.2738156910 -0.6549877  0.002022226  0.2017099245
V3  0.3092612760  0.4903438 -0.721574527  0.5899012750
V4 -0.4608030689 -0.6942426  0.664581293 -0.6923833843
V5  0.4512856042  0.8097183 -0.758953269  0.7489091931
V6  1.0000000000  0.5957745 -0.417471190 -0.0009835832
V7  0.5957745492  1.0000000 -0.524892653  0.4024115144
V8 -0.4174711903 -0.5248927  1.000000000 -0.7722711435
V9 -0.0009835832  0.4024115 -0.772271143  1.0000000000
```

Z macierzy tej wynika, że argument V5 jest wysoce skorelowany z argumentami: V4 oraz V7. Oznacza to, że może on być pominięty w operacji grupowania.


V1
V2
V3
V4
V6
V7
V8
V9
-5 5 20
-40 -10 20
-20 20
-10 20
-20 10
-5 15
0 40
-40 0
-20 20
-20 20
-10 20
-20 0 20
0 40
-20 20
-40 -10 20