Zaawansowane Metody Klasyfikacji i Analiza Skupień

Eksploracja danych - Lista nr 4

Ksawery Józefowski, 277513

2025-06-14

Spis treści

1 Zadanie nr 1

1.1 Wprowadzenie

Celem analizy jest zastosowanie i ocena skuteczności zaawansowanych metod klasyfikacji `random forest` i `bagging`, jak i test algorytmu `SVM` do budowy klasyfikatorów bazujących na różnych jądrach.

1.2 Rodziny klasyfikatorów/uczenie zespołowe

Tabela 1: Macierz pomyłek Bagging

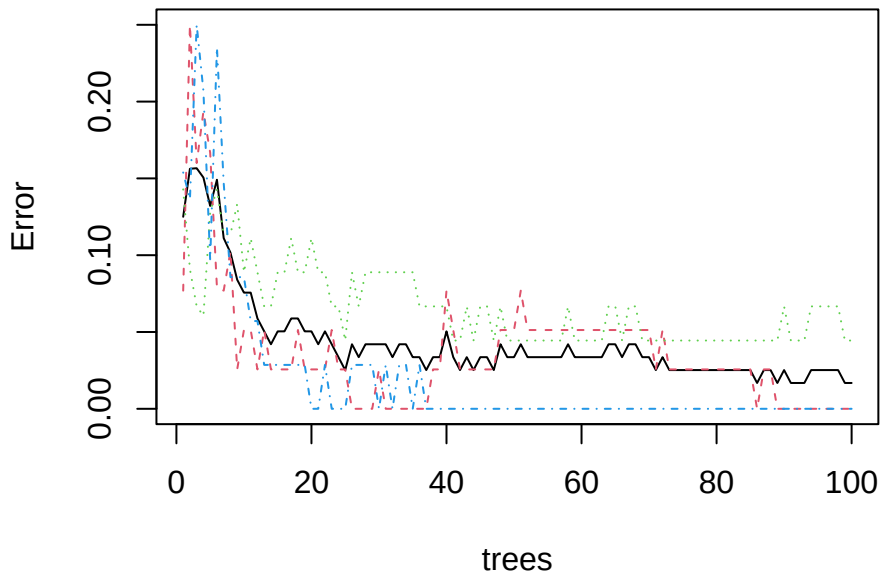
1	2	3
20	0	0
1	25	0
0	1	12

```
## [1] "Błąd klasyfikacji Bagging: 0.034"
```

Tabela 2: Macierz pomyłek Random Forest

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Błąd klasyfikacji Random Forest: 0"
```



Rysunek 1: Wykres błędu Random Forest

```
## [1] "Błąd drzewa: 0.121"
## [1] "Błąd baggingu: 0.065"
## [1] "Błąd random forest: 0.02"
## [1] "Redukcja błędu (bagging): 46.15 %"
## [1] "Redukcja błędu (random forest): 83.13 %"
```

W przeprowadzonym porównaniu metod klasyfikacyjnych na zbiorze danych `wine` zastosowano trzy podejścia: pojedyncze drzewo decyzyjne jako klasyfikator bazowy, oraz dwie metody zespołowe, `bagging` oraz `random forest`. Wyniki klasyfikacji wskazują jednoznacznie, że metody zespołowe pozwalają na istotne zwiększenie dokładności predykcji. Dla pojedynczego drzewa decyzyjnego błąd klasyfikacji wyniósł około `12.1%`. Po zastosowaniu algorytmu `bagging` błąd ten spadł do około `6.5%`, co oznacza względną redukcję błędu klasyfikacji o ponad `46%`. Jeszcze większą poprawę zaobserwowano przy użyciu `random forest`, błąd klasyfikacji spadł do około `2%`, co odpowiada redukcji błędu o ponad `83%` w stosunku do klasyfikatora bazowego.

Zarówno w oparciu o bezpośrednią ocenę klasyfikacji na zbiorze testowym (macierz pomyłek), jak i na podstawie bardziej zaawansowanej metody walidacyjnej `.632+`, metody zespołowe wykazały się wyraźnie lepszą skutecznością. `Random forest` przewyższył skutecznością `bagging`, co wynika z faktu, że dodatkowe losowanie zmiennych na etapie konstrukcji pojedynczych drzew wprowadza większą różnorodność modeli bazowych, a tym samym zwiększa ogólną zdolność klasyfikacyjną zespołu.

Można więc jednoznacznie stwierdzić, że zastosowanie klasyfikatorów zespołowych prowadzi do *istotnej* redukcji błędu klasyfikacji w porównaniu do pojedynczego drzewa decyzyjnego.

1.3 Metoda wektorów nośnych (SVM)

Porównamy skuteczność modeli SVM z różnymi funkcjami jądrowymi:

- Jądro liniowe (kernel="linear")
- Jądro wielomianowe (kernel="polynomial")
- Jądro radialne (RBF) (kernel="radial")

1.3.1 Jądro liniowe z różnymi wartościami `C`

Tabela 3: Macierz pomyłek SVM jądro liniowe (C=0.1)

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Dokładność (C=0.1): 1"
```

Tabela 4: Macierz pomyłek SVM jądro liniowe (C=1)

1	2	3
20	0	0
0	26	0
0	1	12

```
## [1] "Dokładność (C=1): 0.983050847457627"
```

Tabela 5: Macierz pomyłek SVM jądro liniowe (C=10)

1	2	3
20	0	0
0	26	0
0	1	12

```
## [1] "Dokładność (C=10): 0.983050847457627"
```

Na podstawie uzyskanych wyników można zaobserwować wyraźny wpływ wyboru wartości parametru `C` na skuteczność klasyfikacji metodą `SVM`. W przypadku jądra liniowego widzimy, że zmiana parametru `C` ma istotne znaczenie dla dokładności modelu. Dla niskiej wartości `C=0.1` osiągnęliśmy perfekcyjną dokładność na poziomie `100%`, podczas gdy zwiększenie wartości `C` do 1 i 10 spowodowało nieznaczny spadek dokładności do około `98.3%`. Sugeruje to, że w tym konkretnym przypadku mniejsza wartość parametru `C`, odpowiadająca szerszemu marginesowi decyzyjnemu, lepiej radzi sobie z ogólnymi właściwościami danych.

1.3.2 Porównanie różnych jąder

Tabela 6: Macierz pomyłek: SVM jądro wielomianowe (stopień 2)

1	2	3
19	1	0
0	25	1
1	0	12

```
## [1] "Dokładność (wielomian stopnia 2): 0.949152542372881"
```

Tabela 7: Macierz pomyłek: SVM jądro radialne (RBF)

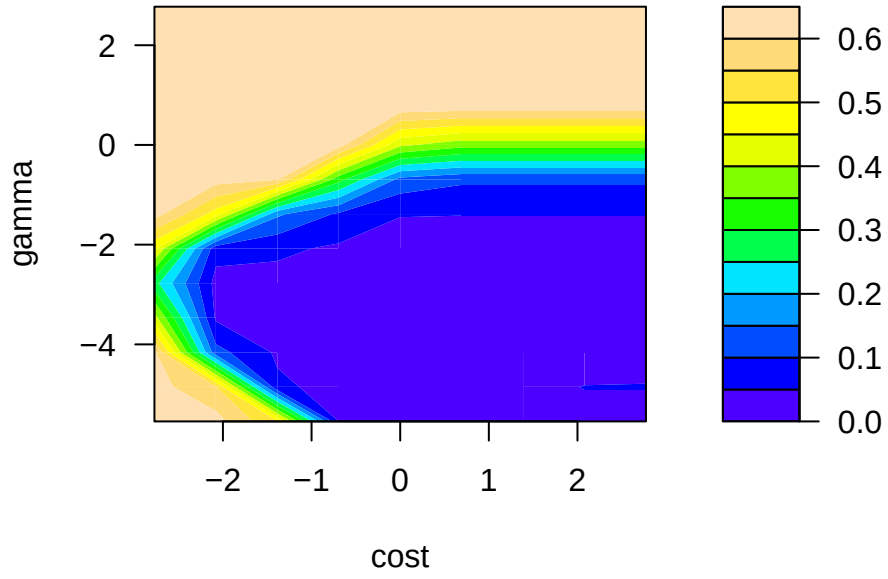
1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Dokładność (radialne RBF): 1"
```

Porównując różne funkcje jądrowe, wyraźnie widać różnice w osiąganej dokładności. Jądro liniowe i radialne dały najlepsze wyniki (`100%` dokładności), podczas gdy jądro wielomianowe stopnia drugiego osiągnęło nieco niższą dokładność na poziomie około `94.9%`. Ta różnica pokazuje, że wybór funkcji jądrowej ma kluczowe znaczenie dla skuteczności klasyfikacji. W tym przypadku zarówno proste jądro liniowe, jak i bardziej złożone jądro radialne doskonale poradziły sobie z separacją klas, podczas gdy jądro wielomianowe okazało się nieco mniej efektywne.

Warto zauważyć, że doskonałe wyniki osiągnięte przez jądro liniowe mogą sugerować, że klasy w zbiorze danych `wine` są liniowo separowalne lub prawie liniowo separowalne w przestrzeni cech. Fakt, że jądro radialne również osiągnęło `100%` dokładność, potwierdza jego elastyczność i zdolność do dopasowania się do różnych struktur danych, chociaż w tym konkretnym przypadku nie było to konieczne, skoro prostsze jądro liniowe dało równie dobre wyniki.

1.3.3 Dostrojone jądro radialne



Rysunek 2: Wydajność SVM

```
##    cost     gamma
## 15    2 0.0078125
```

Tabela 8: Macierz pomyłek: SVM jądro radialne (dostrojone)

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Dokładność (domyślne RBF): 1"
```

```
## [1] "Dokładność (dostrojone RBF): 1"
```

Na podstawie przedstawionych wyników można stwierdzić, że zarówno model SVM z domyślnymi parametrami, jak i model z optymalnie dobranymi parametrami osiągnęły identyczną, perfekcyjną dokładność klasyfikacji na poziomie 100%. Macierze pomyłek dla obu wersji modelu pokazują, że wszystkie próbki ze zbioru testowego zostały poprawnie zaklasyfikowane, bez żadnych błędów.

Proces strojenia parametrów, mimo że zidentyfikował optymalne wartości parametrów (`cost` = 2, `gamma` = 0.0078125), nie przyniósł poprawy w skuteczności klasyfikacji w porównaniu z modelem wykorzystującym domyślne ustawienia. Wykres procesu strojenia sugeruje, że obszar optymalnych parametrów jest stosunkowo płaski, co oznacza, że podobną dokładność można osiągnąć przy różnych kombinacjach wartości `cost` i `gamma`.

Wnioskując, w tym konkretnym przypadku optymalizacja parametrów nie poprawiła skuteczności klasyfikatora, ponieważ model z domyślnymi ustawieniami już osiągał maksymalną możliwą dokładność. Może to wynikać z faktu, że zbiór danych wine jest stosunkowo dobrze separowalny, a proste modele radzą sobie z nim doskonale nawet bez specjalnego dostrajania parametrów.

1.4 Wniosek

Porównując skuteczność metod z punktu `a)` i `b)`, można stwierdzić, że zarówno metody zespołowe (`bagging` i `random forest`), jak i `SVM` osiągnęły bardzo wysoką dokładność klasyfikacji. `Random forest` bazując na metodzie walidacyjnej `.632+` uzyskał prawie perfekcyjny wynik. `SVM` z jądrem liniowym i radialnym sklasyfikował wszystkie próbki poprawnie. `Bagging` wypadł nieco gorzej, ale i tak znacznie lepiej niż pojedyncze drzewo decyzyjne.

W przypadku `SVM` widoczny był wpływ wyboru jądra i parametru `C` na skuteczność, jądro liniowe i radialne okazały się lepsze niż wielomianowe, a niższe wartości `C` w niektórych przypadkach dawały lepsze wyniki. Strojenie parametrów dla jądra radialnego nie poprawiło wyników, ponieważ już domyślne ustawienia dawały perfekcyjną klasyfikację.

Podsumowując, `SVM` z odpowiednim jądrem to najskuteczniejszae i najlepiej dopasowana metoda klasyfikacji dla tego zbioru danych.

2 Zadanie 2

2.1 Wprowadzenie

Celem zadania jest zastosowanie algorytmów analizy skupień - PAM i AGNES oraz ocena i porównanie jakości grupowania.

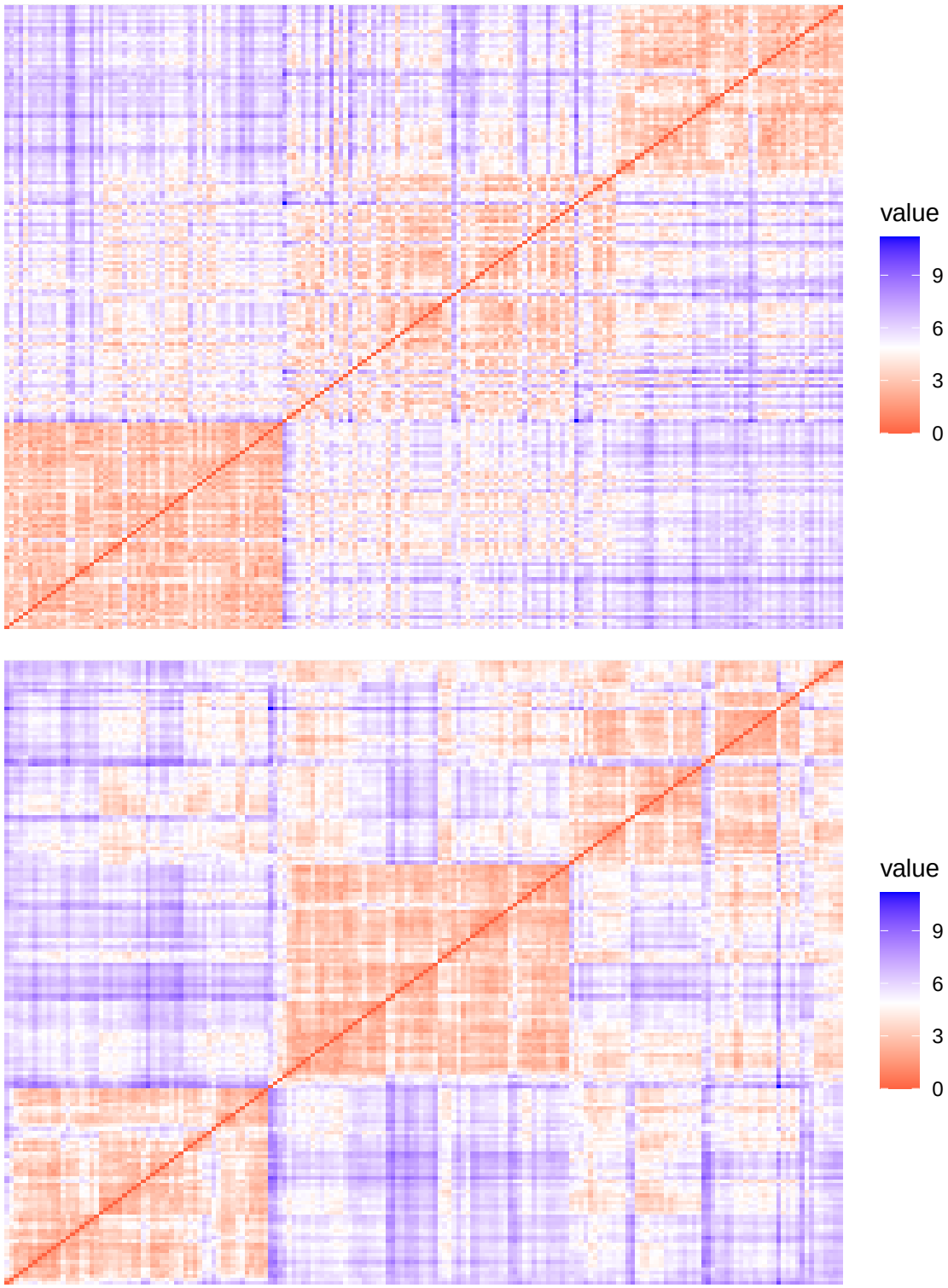
2.2 Wybór i przygotowanie danych

Do analizy wybieramy zbiór `wine`, który zawiera 178 rekordów. Jest to liczba mniejsza niż 200, więc nie będzie potrzeby tworzenia podzbioru. W zbiorze znajdują się zmienne z małymi i dużymi zakresami wartościowymi, więc zostanie zastosowana standaryzacja danych.

Tabela 9: Head zbioru wine

class	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	14	2	2	16	127	3	3	0	2	6	1	4	1065
1	13	2	2	11	100	3	3	0	1	4	1	3	1050
1	13	2	3	19	101	3	3	0	3	6	1	3	1185
1	14	2	2	17	113	4	3	0	2	8	1	3	1480
1	13	3	3	21	118	3	3	0	2	4	1	3	735
1	14	2	2	15	112	3	3	0	2	7	1	3	1450

2.3 Macierz odmienności



Rysunek 3: Macierze odmienności, odpowiednio order FALSE i order TRUE

2.4 Metoda PAM

Rysunek 4: Skupienia metodą PAM na przestrzeni MDS

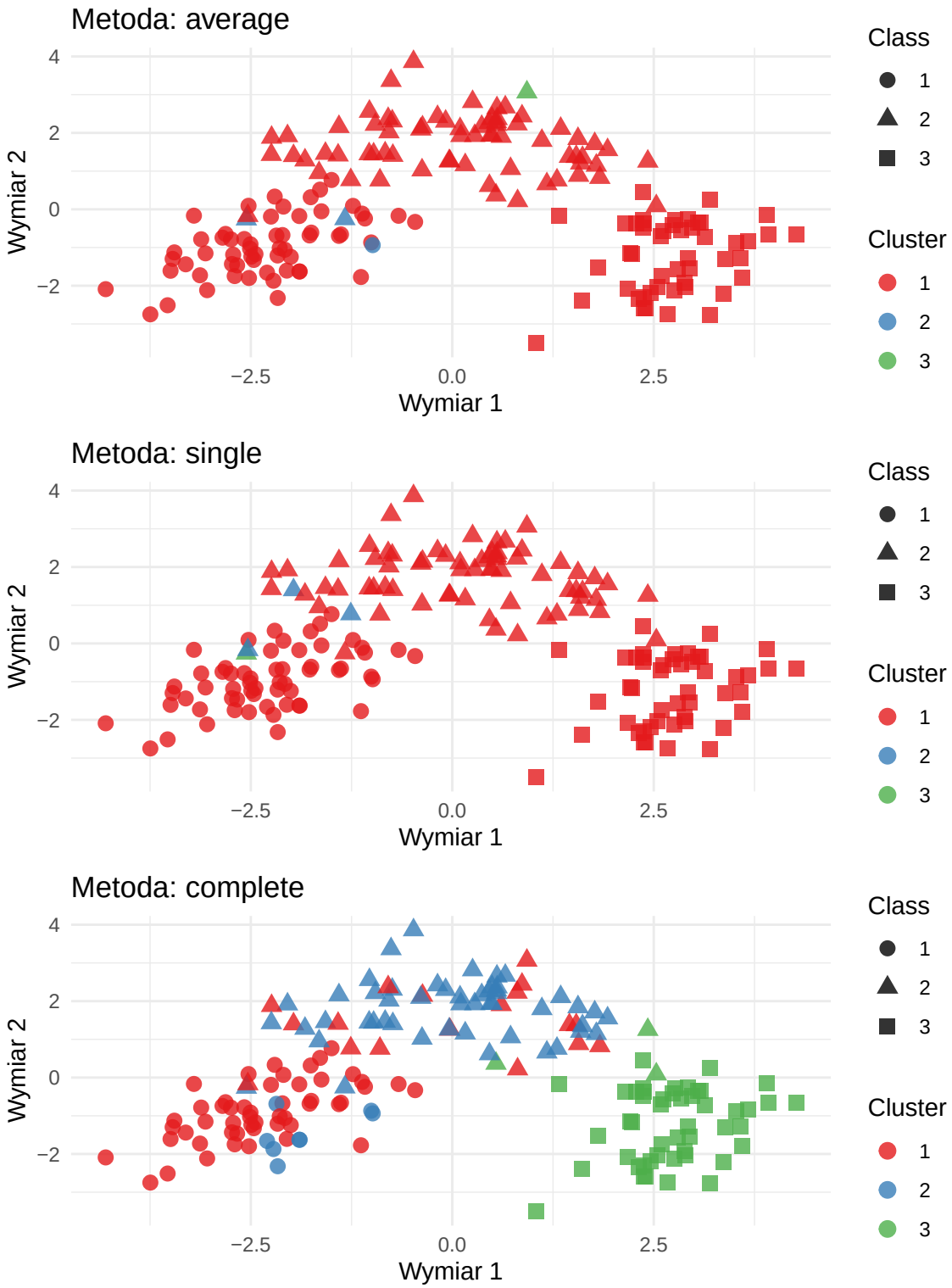
W tym przypadku zastosowano metodę `MDS`, aby przedstawić dane wielowymiarowe na dwuwymiarowym wykresie rozrzutu. Ponieważ dane po skalowaniu mają aż *13* cech liczbowych, bez redukcji wymiaru ich bezpośrednia wizualizacja byłaby niemożliwa. W tym przypadku użycie `MDS` jest możliwe, ponieważ dane są ciągłe, numeryczne i przeskalowane, dzięki czemu odległość euklidesowa dobrze oddaje podobieństwa między obiektami. Zbiór zawiera stosunkowo niewielką liczbę obserwacji (*178*), co pozwala `MDS` efektywnie odwzorować relacje między nimi bez dużej utraty informacji.

Na podstawie wykresu przedstawiającego wyniki grupowania metodą `PAM` w przestrzeni dwuwymiarowej, można sformułować kilka istotnych obserwacji dotyczących własności otrzymanych skupień oraz ich zgodności z rzeczywistym podziałem na klasy.

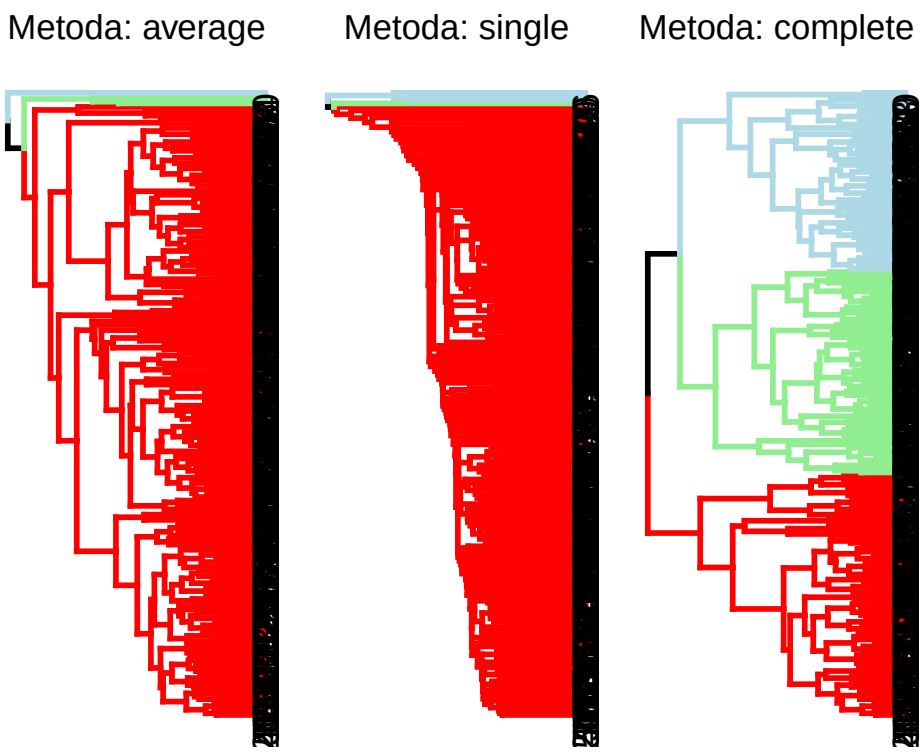
Po pierwsze, skupienia wykazują bardzo dobrą separację. Obiekty należące do różnych klastrów tworzą wyraźnie odseparowane grupy w przestrzeni dwuwymiarowej. Szczególnie widoczna jest granica między klastrem czerwonym i zielonym. Taka wyraźna separacja wskazuje, że metoda `PAM` skutecznie rozpoznała strukturę skupień w danych. Dodatkowo skupienia są zwarte, obiekty w ramach każdego klastra są do siebie blisko rozmieszczone, co świadczy o ich jednorodności i wewnętrznej spójności.

Jeśli chodzi o zgodność uzyskanych skupień z rzeczywistym podziałem na klasy, to jest ona dobra, ale nie idealna. Klasa 1 w całości odpowiada klastrowi czerwonymu, klasa 2 jest w większości przypisana do klastra niebieskiego, ale zauważalne są obserwacje odstające przypisane, też do innych klastrów. Klasa 3 w całości została przypisana do klastra zielonego. Oznacza to, że metoda `PAM` przy liczbie klastrów równej trzy dobrze odwzorowała rzeczywisty podział na klasy 1 i 3, lecz miała problemy z klasą 2. Sugeruje to, że algorytm nie był w stanie jednoznacznie przypisać tej klasy do jednego skupienia.

2.5 Algorytm AGNES



Rysunek 5: Skupienia algorytmem AGNES na przestrzeni MDS dla różnych metod



Rysunek 6: Dendogramy różnych metod AGNES

Podobnie jak w poprzednim przypadku została zastosowana `MDS` do redukcji wymiaru. Analizując zarówno wykresy, jak i dendrogramy przedstawiające wyniki grupowania hierarchicznego algorytmem `AGNES` dla metod `average`, `single` oraz `complete` na zbiorze danych `wine`, można zauważyć istotne różnice w podstawowych własnościach otrzymanych skupień.

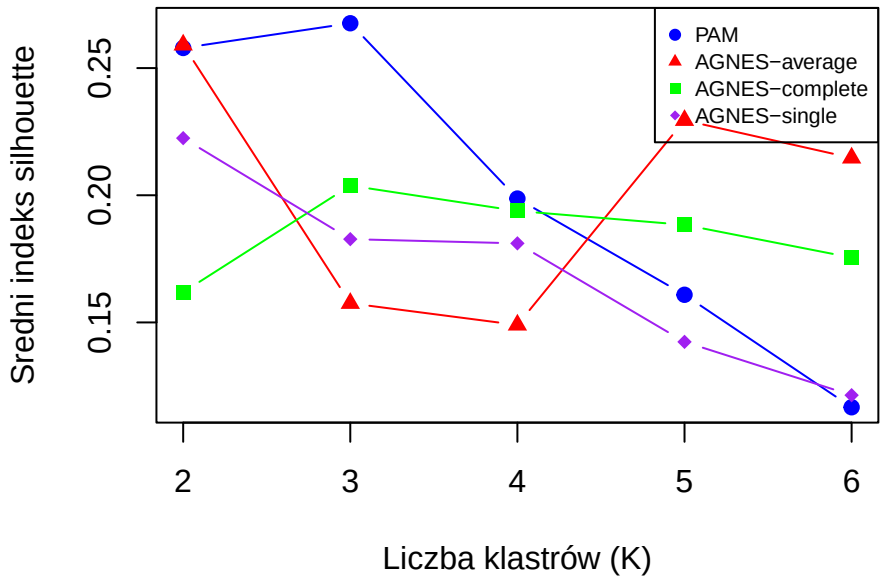
Metoda `average` polega na obliczaniu średniej odległości między wszystkimi parami punktów należącymi do różnych skupień. Na wykresach i dendrogramach widać, że większość obiektów została przypisana do jednego dużego klastra, przez co jednorodność i separacja między klastrami są ograniczone. Powoduje to niewielką zgodność wynikowych klastrów z rzeczywistą przynależnością obiektów do klas, co oznacza, że ta metoda nie odtwarza w pełni struktury klas obecnych w danych.

Metoda `single` definiuje odległość pomiędzy skupieniami jako najmniejszą odległość między dowolną parą punktów, po jednej z każdego skupienia. Skutkuje to efektem „łańcuchowania", czyli łączeniem skupień jedno po drugim na podstawie pojedynczych bliskich sobie punktów. Dendrogram dla tej metody jest najbardziej rozciągnięty, a wykresy pokazują, że niemal wszystkie obiekty również trafiają do jednego dużego klastra. Takie skupienia są mało zwarte, mają rozmyte granice, są bardzo wrażliwe na pojedyncze punkty i nie odpowiadają dobrze rzeczywistej strukturze klas.

Metoda `complete` polega na braniu pod uwagę maksymalnej odległości pomiędzy punktami z różnych skupień. Dzięki temu uzyskane skupienia są wyraźnie bardziej zwarte, jednorodne i lepiej odseparowane od siebie, co można zobaczyć zarówno na dendrogramach, jak i na wykresach `MDS`. Gałęzie na dendrogramie są bardziej zwarte, a na wykresach grupy są lepiej oddzielone i bardziej jednorodne pod względem przypisania do rzeczywistych klas. Choć także w przypadku tej metody nie uzyskuje się idealnego pokrycia klastrów z rzeczywistymi klasami, to jednak podział jest zdecydowanie bardziej zgodny z rzeczywistą strukturą zbioru `wine`.

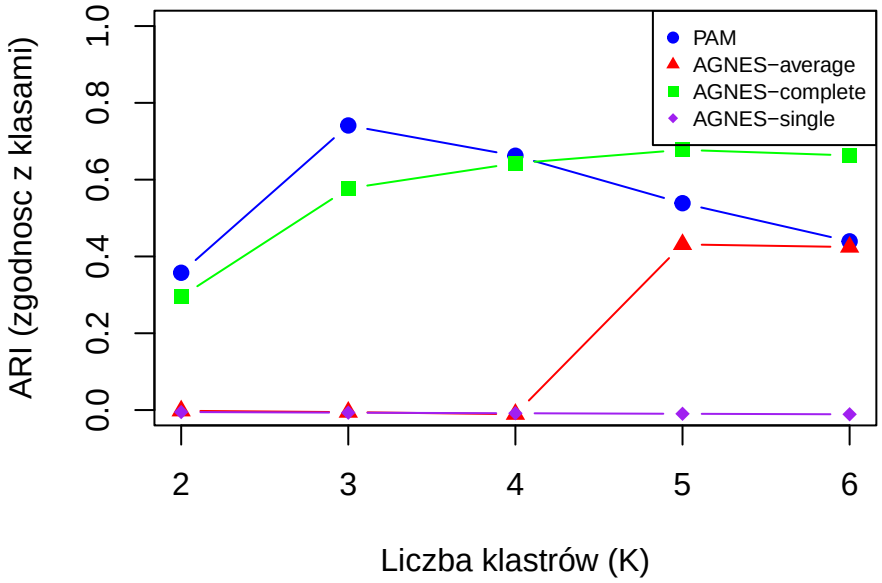
2.6 Ocena jakości grupowania

2.6.1 Wskaźniki wewnętrzne



Rysunek 7: Wartość silhouette dla PAM i AGNES

2.6.2 Wskaźniki zewnętrzne



Rysunek 8: Zgodność grupowania z klasami

Na podstawie przedstawionych wykresów można stwierdzić, że najlepsze rezultaty grupowania dla zbioru danych `wine` uzyskano, stosując algorytm `PAM` lub algorytm `AGNES` z metodą complete.

W przypadku wskaźnika silhouette `Rysunek 7`, który mierzy zwartość i separację skupień, najwyższe wartości osiągnięto dla algorytmu `PAM` przy liczbie klastrów równej *3*. Wartość silhouette dla pozostałych metod jest niższa niezależnie od liczby klastrów, co świadczy o mniejszej czytelności i zwartości otrzymanych podziałów.

Drugim istotnym wskaźnikiem jest `ARI` (Adjusted Rand Index), który przedstawia zgodność grupowania z rzeczywistym podziałem na klasy `Rysunek 8`. Najwyższe wartości `ARI` obserwowane są dla `PAM` oraz `AGNES` complete przy odpowiednio liczbie klastrów *3* dla `PAM` oraz *5* dla `AGNES`, co oznacza, że te algorytmy najlepiej odwzorowują rzeczywistą strukturę danych. W szczególności algorytm `PAM` osiąga bardzo wysoką wartość `ARI` przy *3*, a metoda `AGNES` complete daje rezultaty tylko nieznacznie gorsze. Metody `AGNES` average i `AGNES` single charakteryzują się bardzo niską zgodnością niezależnie od przyjętego `K` i nie odwzorowują poprawnie rzeczywistego podziału na klasy.

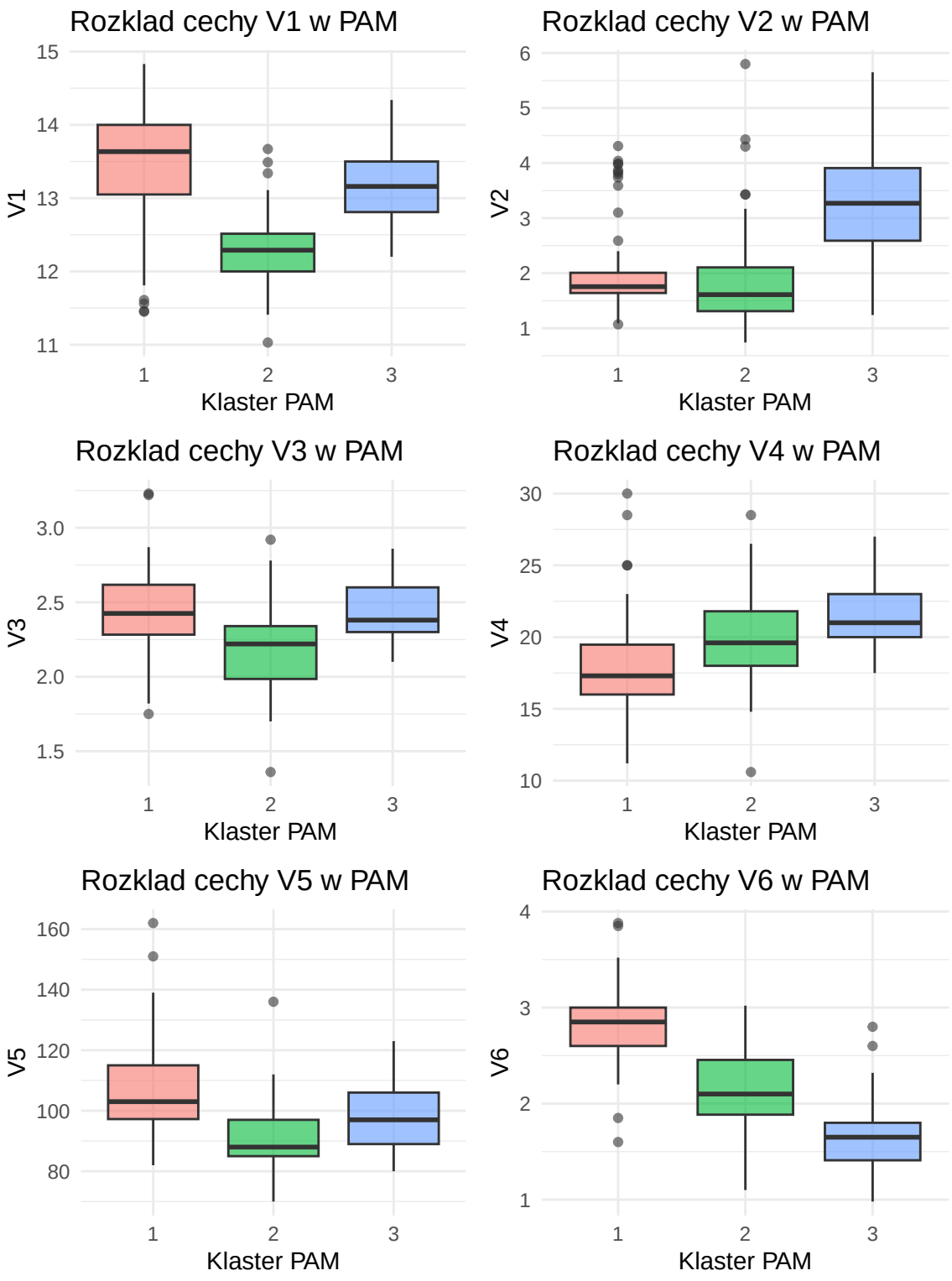
2.7 Interpretacja wyników grupowania

Tabela 10: Średnie wartości cech w klastrach PAM (K=3)

Klaster	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	13	2	2	18	107	3	3	0	2	5	1	3	1018
2	12	2	2	20	91	2	2	0	1	3	1	3	489
3	13	3	2	21	99	2	1	0	1	7	1	2	628

Tabela 11: Średnie wartości cech w klastrach AGNES-complete (K=3)

Klaster	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	13	2	2	17	105	3	3	0	2	5	1	3	985
2	12	2	2	21	94	2	2	0	2	3	1	3	573
3	13	3	2	21	99	2	1	0	1	7	1	2	622



Rysunek 9: Rozkłady cech V1-V6 w klastrach (PAM)

Na podstawie wskaźników oceny jakości grupowania ustalono, że optymalną liczbą skupień w analizowanym zbiorze wine jest *3*, a najlepsze rezultaty uzyskano stosując algorytm `PAM`.

Analizując średnie wartości cech dla poszczególnych klastrów (`tabela 10` dla `PAM` oraz `tabela 11` dla `AGNES`), można zauważyć, że każda z wyodrębnionych grup istotnie różni się pod względem przynajmniej kilku zmiennych opisujących próbki wina. Przykładowo:

- Klaster *1* w obu metodach charakteryzuje się wyższymi średnimi wartościami cechy `V1`, `V3`, `V4` oraz `V5`, co może świadczyć o wyższej zawartości danego składnika lub intensywności konkretnej właściwości produktu w tym skupieniu.
- Klaster *2* wyróżnia się relatywnie niższymi wartościami większości cech, zwłaszcza `V1`, `V3`, `V5` i `V13`, co potwierdzają zarówno zestawienia średnich, jak i wykresy pudełkowe. Na wykresach wyraźnie widać, że obiekty w tym klastrze mają niższe wartości np. dla cech `V1` i `V5`.
- Klaster *3* prezentuje natomiast nieco wyższe wartości cech `V1`, `V4` i `V13` w porównaniu do klastra *2*, lecz niższe niż klaster *1*, co świadczy o pośrednim "profilu" tej grupy.

Wykresy pudełkowe dla cech `V1`–`V6` pozwalają jeszcze lepiej zobrazować te różnice, zaobserwować można wyraźne oddzielenie rozkładów poszczególnych cech dla różnych klastrów, zwłaszcza pod względem poziomu i rozrzutu wartości, co potwierdza trafność dokonanego podziału i sugestię, że skupienia są pod względem tych parametrów wewnętrznie jednorodne, a zróżnicowane między sobą.

Tabela 12: Obiekty-medoidy w PAM wraz z wartościami cech

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	Cluster_PAM
36	13	2	2	20	100	3	3	0	2	5	1	3	920	1
107	12	2	2	19	80	2	2	0	2	3	1	3	510	2
149	13	3	2	22	92	2	1	0	1	8	1	2	650	3

W przypadku metody `PAM` istotnym elementem interpretacji jest identyfikacja medoidów, czyli najbardziej reprezentatywnych obserwacji dla każdego klastra (`tabela 12`). Dla każdej z trzech grup wskazano konkretny obiekt, który najlepiej odzwierciedla typowy profil cech danej grupy. Analizując wartości cech tych medoidów, można zobaczyć, że:

- Medoid klastra *1* charakteryzuje się najwyższymi wartościami cech `V1`, `V5` i `V13`, co jest spójne z ogólną charakterystyką tego skupienia i potwierdza obecność w tej grupie "najbogatszych" pod względem tych cech win.
- Medoid klastra *2* odznacza się najniższymi wartościami większości cech, przede wszystkim `V5` i `V13`, co czyni go typowym reprezentantem grupy mniej "intensywnej" kompozycyjnie.
- Medoid klastra *3* prezentuje wartości pośrednie z wyraźnie podwyższoną cechą `V4` i relatywnie wysokim `V13`.

2.8 Podsumowanie

Podsumowując, przeprowadzona analiza pozwala uznać, że zarówno algorytm `PAM`, jak i `AGNES` z metodą complete skutecznie oddzieliły od siebie grupy próbek wina charakteryzujące się różnym poziomem wybranych cech chemicznych. Jednocześnie, wskazane medoidy w metodzie `PAM` dobrze reprezentują typowe właściwości swoich klastrów i dodatkowo ułatwiają interpretację, pozwalają wskazać "wzorcowe" obserwacje, które stanowią punkt odniesienia dla danej grupy. Całość potwierdza wysoką spójność wewnętrzną oraz wyraźne zróżnicowanie między klastrami w wyselekcjonowanych wymiarach.