Indukcyjna Analiza Danych
Laboratorium 3
Klasteryzacja

Piotr Błoński

7 maja 2020

Prowadzący: Dr inż. Paweł Myszkowski

1 Cel ćwiczenia

Celem zadania laboratoryjnego jest dalsze zapoznanie się z językiem R i poznanie uczenia nienadzorowanego na przykładzie zadania klasteryzacji.

2 Wstęp

Przed podejściem do wykonywania zadania laboratoryjnego należało zapoznać się z zagadnieniami teoretycznymi i odpowiedzieć na pytania pomocnicze.
Podstawowe pojęcia:

a. Klaster - podzbiór danych o podobnych cechach. Potencjalnie o tej samej klasie.

b. Medoid - jest to obiekt ze zbioru którego średnia odmienność od innych objektów w klastrze jest minimalna.[4]

c. Centroid - Centroid jest to reprezentant danego skupienia lub inaczej środek danej grupy.

d. Manhattan distance - tą miarę odległości oblicza się na zasadzie $distance = |x_1 - x_2| + |y_1 - y_2| + ...$

2.1 Teoria - k-means

Jest to jeden z podstawowych algorytmów analizy skupień. Celem algorytmu jest przypisanie do wektorów kodowych r_i (przy założeniu że $i \in [1, N]$) M n-wymiarowych wektorów danych, przy jak najmniejszym średnim błędzie kwantyzacji określony wzorem $D = \frac{1}{K} * \sum_{i=1}^{K} d*(x_i, r)$ gdzie K jest liczbą elementów x_i przypisanych do wektora kodowego r, natomiast d miarą błędu kwantyzacji

i najczęściej jest to błąd kwadratowy określany dla wektorów n-wymiarowych jako: $d(x,r) = \sum_{j=1}^{n}(x_j - r_j)^2$.
Przebieg algorytmu:

a. Wybierz N wektorów kodowych i określ maksymalny błąd kwantyzacji e.

b. $m := 0$ iterator na 0

c. $D_m := \infty$ (średni błąd kwantyzacji w m-tej iteracji czyli zerowej)

d. Dopóki nie uzyskano zadowalającego rezultatu powtarzaj:

   (a) Podziel M wektorów danych na N grup. Gdzie wektor $x_j (j \in [1, M]$ jest przypisywany do i-tej grupy wtedy i tylko wtedy gdy zachodzi nierówność $d(x_j, r_i) <= d(x_j, r_k)$ dla wszystkich r_k różnych od r_i

   (b) Wyznacz średni błąd kwantyzacji $D_m = \frac{1}{M}\sum_{i=1}^{M} d(x_i, r)$, przy czym do obliczeń brany jest wektor kodowy r z tej grupy, do której został zakwalifikowany wektor danych x_i.

   (c) Wyznacz centroidy dla wszystkich i grup wektorów i przypisz je do wektorów kodowych r_i

   (d) Jeśli $\frac{D_m - 1 - D_m}{D_m} < e$ zakończ pętlę.

Algorytm sukcesywanie dopasowuje wektory kodowe do istniejących danych i w miarę potrzeby przesuwa błędnie zakwalifikowane wektory danych do innych grup. Problem stanowi jednak początkowy wybór wektorów kodowych (pierwszy punkt algorytmu).[1]

2.2 Teoria - PAM

PAM czyli (ang. Partitioning Around Medoids) jest kolejnym algorytmem klasteryzacji dla uczenia nienadzorowanego. PAM jest realizacją metody k-medoidowej, czyli techniki grupowania, która dzieli zbiór danych zawierających n obiektów na K grup (dalej jako klastrów). Jego działanie jest zbliżone do powyższego k-means. różni się tym że centroidy zostają zapisane przez medoidy czyli najbardziej centralne obiekty ze zbioru danych, dla których odległość od wszystkich pozostałych elementów wewnątrz danej grupy jest minimalna. Algorytm ten dąży do minimalizacji sumy odległości wszystkich elementów niebędących medoidami od najbliższych im medoidów. Kolejną różnicą jest sposób definiowania dystansu między obserwacjami, PAM używa norm Manhattan zmiast odległości euklidesowej. Zaletą PAM jest jego odporność na obserwacje odstające (ang. outliers) oraz szumy występujące w danych (ang. robustness). [2]
Przebieg algorytmu PAM:

a. Faza budowy:

   (a) Podziel zbiór danych na K skupień z przypisanymi K medoidami

   (b) Oblicz macierz odległości pomiędzy medoidami oraz pozostałymi obserwacjami

(c) Przypisz każdą z obserwacji (nie będącą medoidem) do nabardziej zbliżonego skupienia.

b. Faza zamiany:

(a) Przy użyciu iteracji zastąp jeden z medoidów jednym z niemedoidów i sprawdź, czy odległości wszystki elementów niebędących medoidami od najbliższych im medoidów są mniejsze

(b) Jeśli nastąpiła przynajmniej jedna zmiana medoidów, przejdź do punktu (c). Jeśli nie zakończ działanie.

2.3 Metryki

W celu oceny naszych modeli możemy użyć poniższych metryk:

a. Rand index jest to miara która pozwala określić czy para instacji została przypisana do tego samego klastra. Zakłada się że dane zostały podne klasteryzacji różnymi metodami a następnie w ten sposób porównane. $RI = \frac{a+b}{\binom{n}{2}}$

b. Dunn index Miara ta zdefiniowana jest jako stosunek minimalnej odległości między próbkami należącymi do różnych klastrów, a maksymalnej odległości próbkami w jednym klastrze. Wartość miary powinna być maksymalizowana. $DI = \frac{d_{min}}{d_{max}}$

c. Purity Miara sprawdza czystość klastrów (w jakim stopniu zawierają instancję jednej klasy) $P = \frac{1}{N} \sum_{m \in M} \max_{d \in D} |m \cup d|$

d. Davies–Bouldin index - Miara, która uwzględnia rozrzut próbek wewnątrz klastra oraz odległości między klastrami. Wartość miary powinna być minimalizowana.Ma to tę wadę, że ma dobry stosunek zgłaszane za pomocą tej metody nie oznacza najlepszy wyszukiwania informacji.[3] $DBI = \frac{1}{K} \sum_{k=1}^{K} M_k = \frac{1}{K} \sum_{k+1}^{K} \max_{k` \neq k} (\frac{\delta_k + \delta_{k`}}{\Delta_{k`k}})$

2.4 Odpowiedzi na pytania pomocnicze

- Czy dane muszą być dyskretyzowane i/lub normalizowane?
- Czy kroswalidacja jest potrzebna? - Tak ponieważ może się zdarzyć że różne lepsze jako medoidy.
- Czym różnią się oba algorytmy? - PAM szuka medoidów i chce zminimalizować sume odległości wszystkich elemtów od najbliższych im medoidów, a k-means chce znaleźć takie wektory danych które maja jak najmniejszy średni błąd kwantyzacji.
- Jakie mają parametry?

a. k-means - k liczba klastrów na które chcemy podzielić dane. Można też podać centra.

b. PAM - można podać między innymi Medoidy, i także ilość klastrów,

- Który z algorytmów jest podatniejszy na szum w danych i „outliery"? Dlaczego? - k-means jest bardziej podatny na outlierów ponieważ średnia wartość jest dość podatna na wartości odstające. Tego algorytmu można stosować do wyszukiwania outliarów.

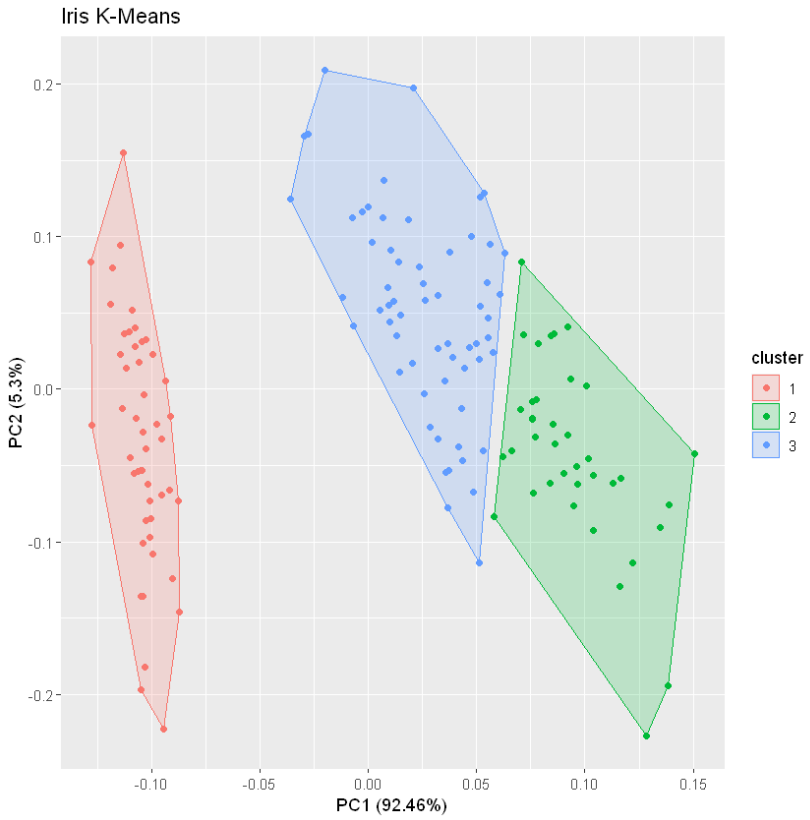
- Czy istnieje potrzeba powtarzania uzyskanych wyników? - PAM - powinien sprawdzić każdą parę medoidów więc nie. - k-means - w zależności gdzie początkowo ustawimy centroidy wyniki mogą być różne.

- Czy sposób mierzenia odległości (miar) wpływa na skuteczność algorytmów? - Tak, w przypadku K-means zamiast odległości euklidesowej możemy użyć inny np. Chebychev Distance. W zależności od danych może mieć to różny wpływ.

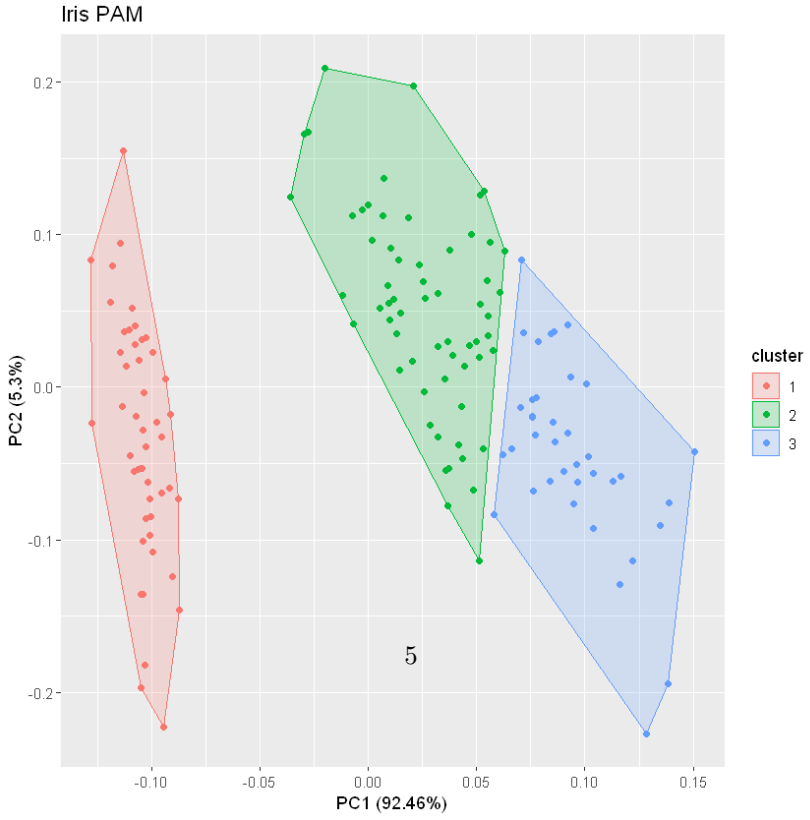
- Co mierzą wskazane miary jakości klasyfikacji i jakie są wartości „optymalne". Np. jakie wartości może przyjąć miara zadana miara (np. DBI) gdy mamy tylko jeden klaster, a jaką wartość jeśli mamy tyle klastrów co instancji (danych)? - wyżej w miarach.

- Jak zinterpretować wartości miary „Purity" dla wybranych modeli? - wyżej w miarach

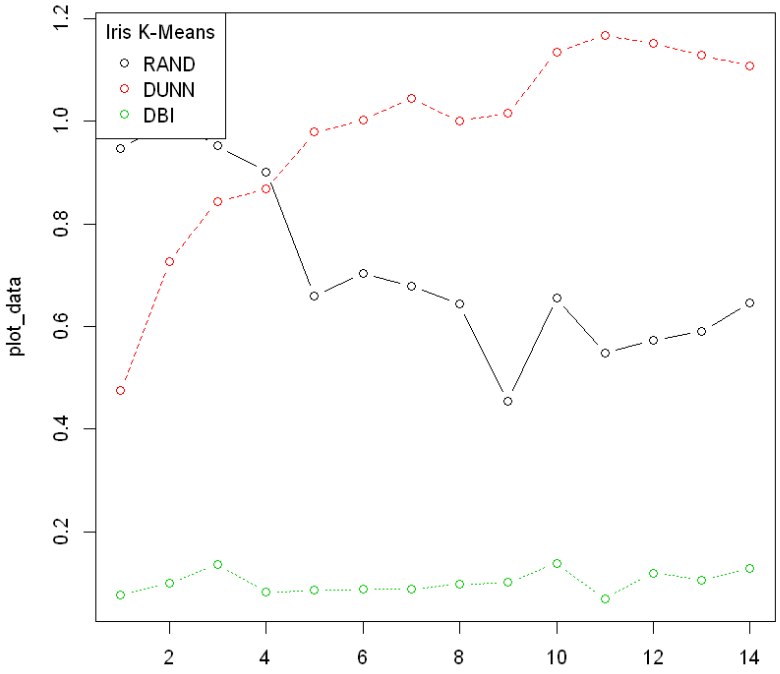
3 Wizualizacja klastrów i wykresy metryk



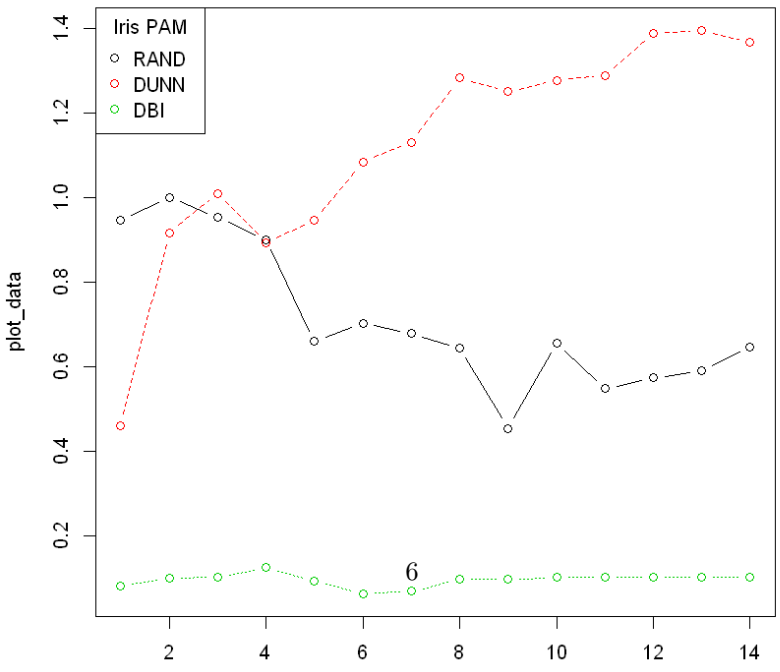
(a) K-means



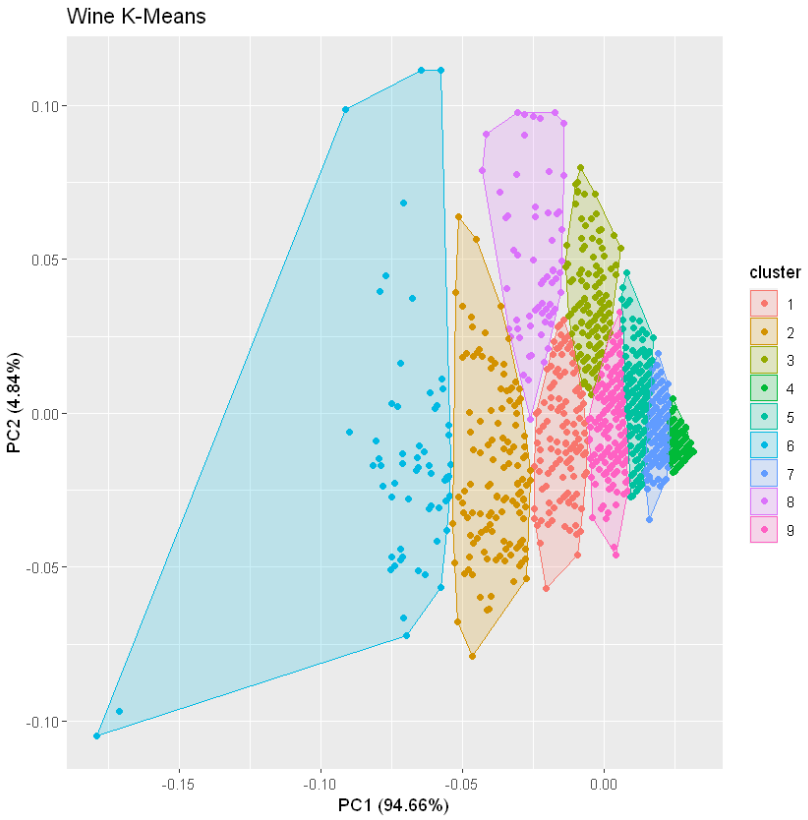
(b) PAM



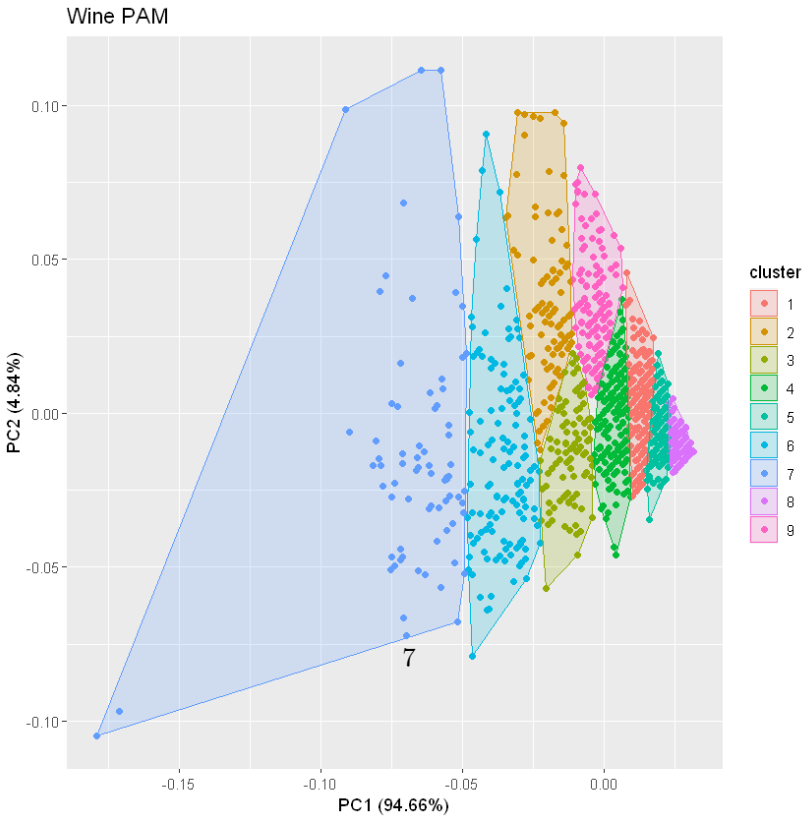
(a) K-means



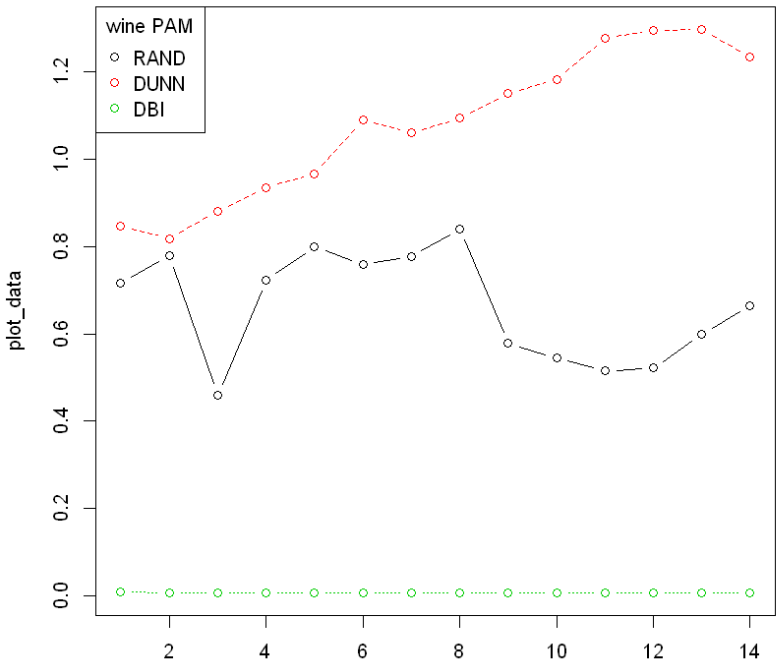
(b) PAM



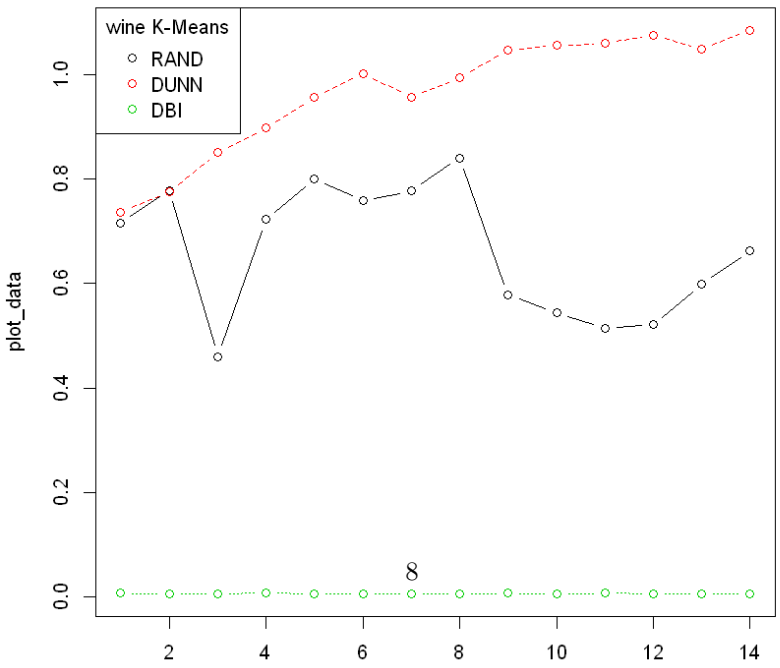
(a) K-means



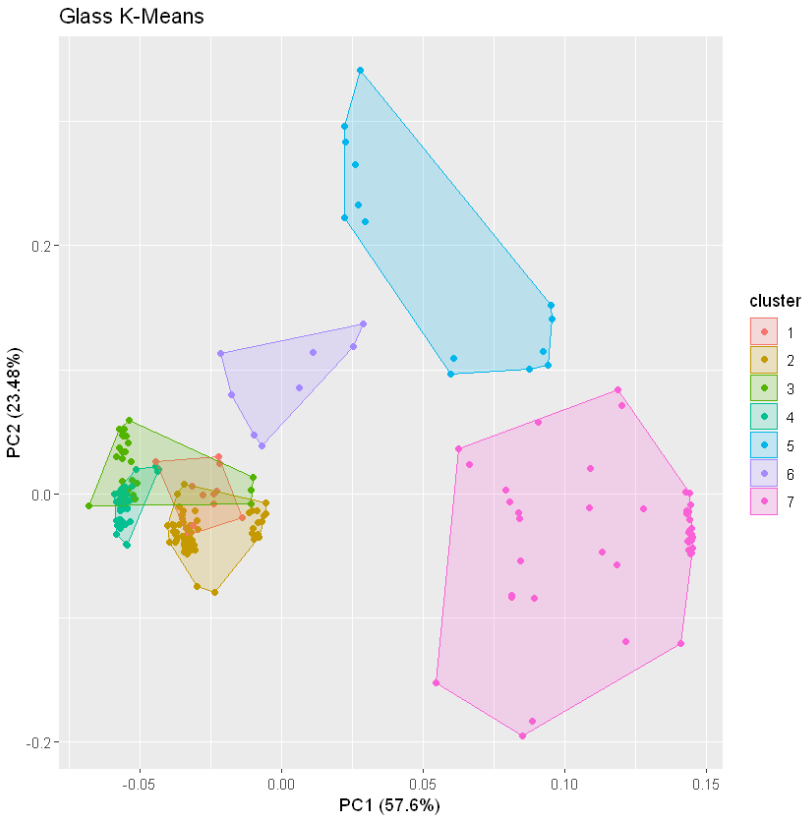
(b) PAM



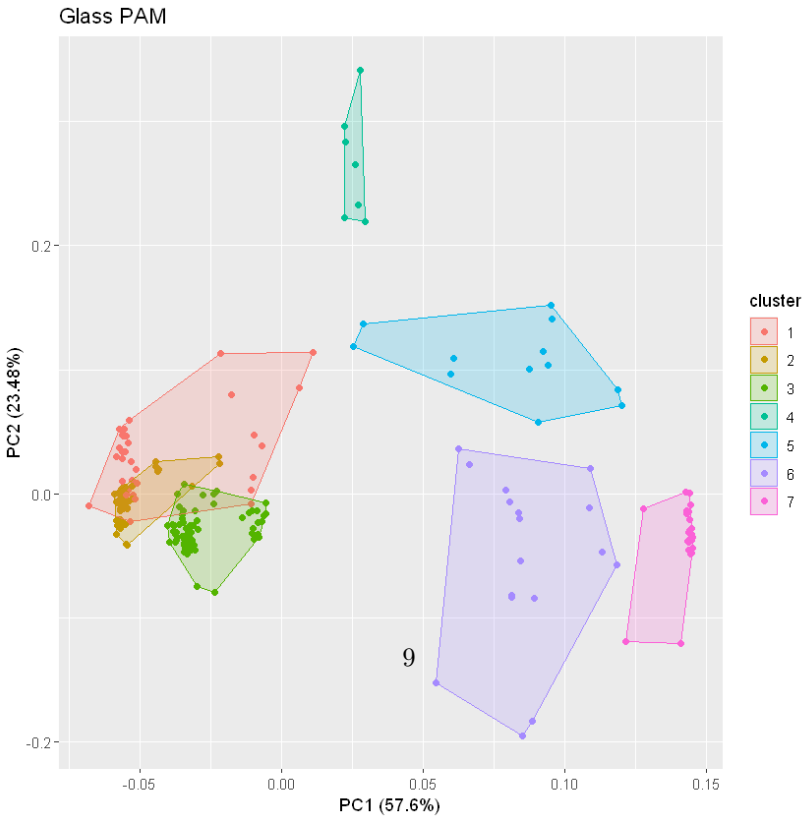
(a) K-means



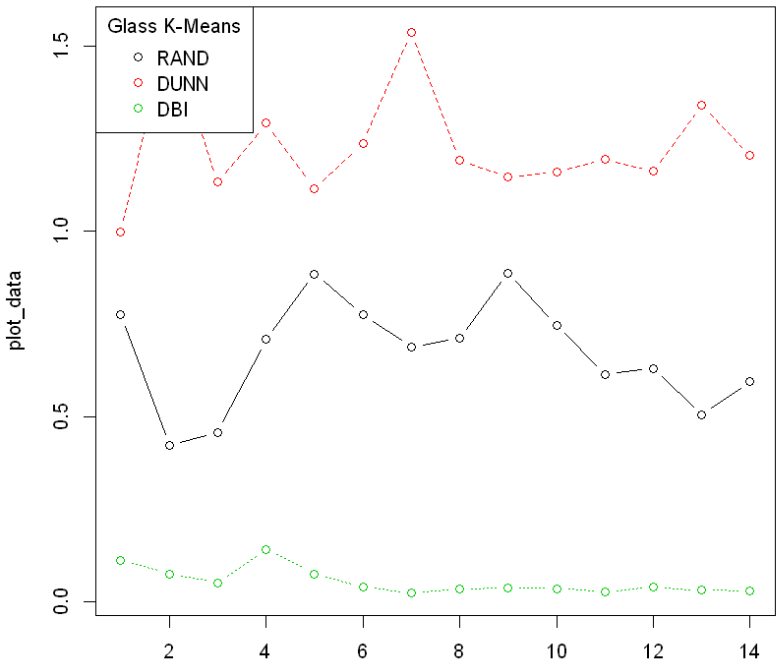
(b) PAM



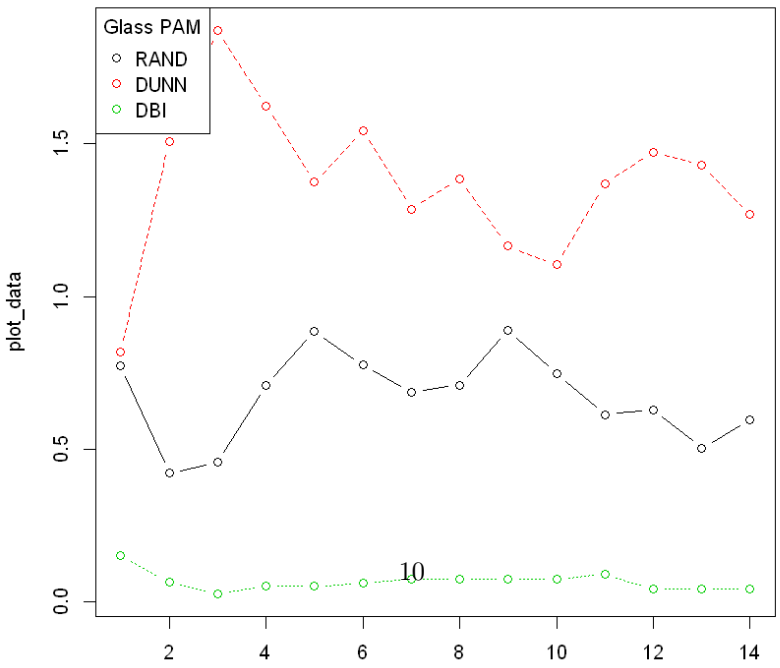
(a) K-means



(b) PAM



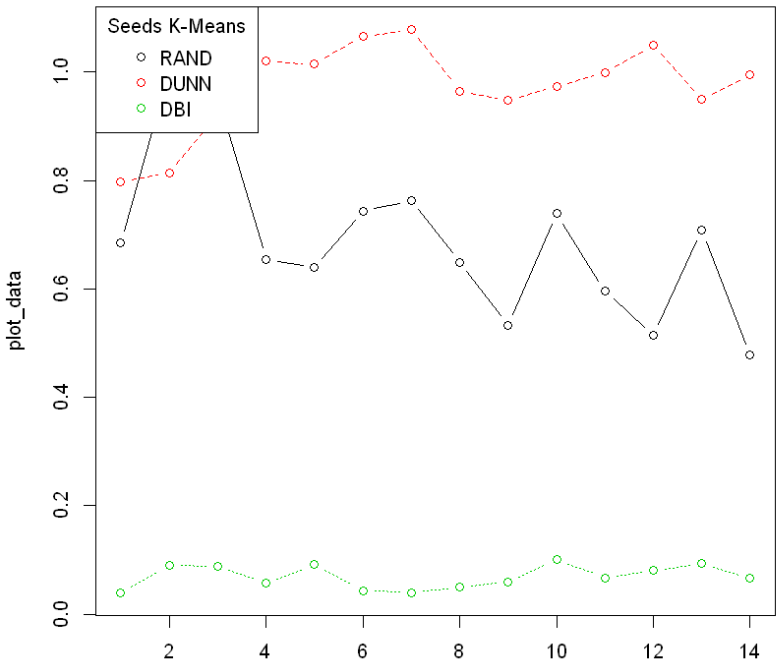
(a) K-means



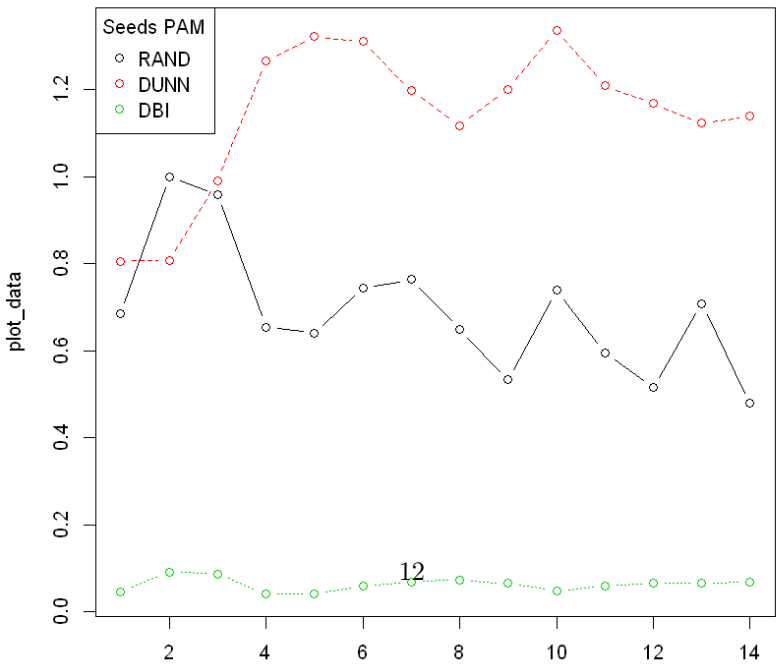
(b) PAM

(a) K-means

(b) PAM



(a) K-means



(b) PAM

4 Wnioski

a. Gotowe algorytmy klasteryzacji są dość proste w użyciu

b. W języku R są gotowe pakiety do wizualizacji klastrów

c. Algorytm K-means może pomóc w znalezień outlierów.

d. Dzięki uczeniu nienadzorowanemu możemy otrzymać zbiory danych.

e. Istnieją metryki które pozwalają porównywać ze sobą klastry

References

[1] *Algorytm centroidów - Kmeans*. URL: https://pl.wikipedia.org/wiki/Algorytm_centroid.

[2] *Algorytm PAM*. URL: https://pl.wikipedia.org/wiki/Algorytm_PAM.

[3] *Davies-Bouldin Index*. URL: https://pl.qwe.wiki/wiki/Davies%E2%80%93Bouldin_index.

[4] *Medoid*. URL: https://en.wikipedia.org/wiki/Medoid.