Sprawozdanie

Indukcyjne Metody Analizy Danych

Ćwiczenie 3. Wybrane metody klasteryzacji w oparciu o system R

Autor: Paweł Mielniczuk

Spis treści:

1. Opis działania algorytmów K-means i PAM

K-means – jest to algorytm z dziedziny algorytmów uczenia nienadzorowanego. Stosowany jest do rozwiązywania problemów klasteryzacji. Za zadanie ma na celu utworzenia *k* klastrów, do których zostaną przydzielone i w ten sposób pogrupowane dane wejściowe.

Algorytm na początku ustawia *k* centroidów losowo rozmieszczonych w różnych lokalizacjach, po czym obliczana jest odległość każdej danej wejściowej od każdego z centroidów. Następnie wybierana jest najmniejsza odległość i w ten sposób przypisywana jest zależność do danego klastra.

Następnie dla każdego klastra wyliczany jest nowy środek. Brane są wszystkie dane, które należą do danego centroida i wyliczana jest średnia z ich wektorów. Każdy atrybut jest sumowany z innymi i dzielony przez liczbę instancji należących do tego centroida. Ta nowa średnia mówi o tym w jakie miejsce przesunie się centroid.

Kroki te powtarzane są dopóki żadne z danych wejściowych nie zmienią przypisania do innego klastra.

K-medoids – jest to algorytm podobny do k-means i jest pewnym jego ulepszeniem. K-means jest wrażliwe na instancje brzegowe, odbiegające mocno od innych. Wartości takie wpływają znacząco na średnią wartość i cechują się tym, że „przyciągają” do siebie średnią i zaburzają rozkład danych. W przypadku algorytmu PAM zamiast średniej używana jest najbardziej centralna instancja klastra, zwana medoidem.

Na początku zamiast wybierać losowo punkty, które tworzą środki naszych klastrów wybierane są losowo instancje, które stają się medoidami. Tak jak w przypadku k-means przydział do klastra dzieje się poprzez wybranie najmniejszej odległości.

Całkowity koszt klasteryzacji obliczany jest jako suma odległości pomiędzy instancjami do ich środka klastra.

Wybranie nowego środka klastra dzieje się poprzez zamianę po kolei każdego z punktów należącego do danego klastra, nie będącego medoidem, z aktualnym medoidem. Jeżeli całkowity koszt klasteryzacji jest mniejszy od poprzedniego wtedy

zamiana była dobrym posunięciem, jeżeli nie zmniejszyła się wtedy odwracana jest zamiana.

Jeżeli żaden medoid nie został zamieniony algorytm uważa się za zakończony.

2. Wprowadzenie do zbiorów danych

Podczas analizy i implementacji użyte zostały cztery zbiory danych. Zbiory podzielone są na dwie części. Pierwszą z nich są cechy, dokładnie wektor, cech oraz etykiety mówiące o przynależności wektora cech do konkretnej klasy.

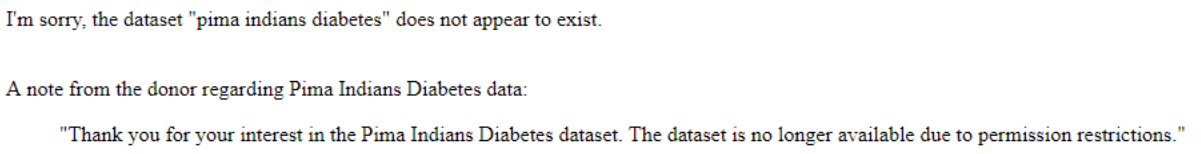
Wszystkie zbiory dostępne są do pobrania ze strony
https://archive.ics.uci.edu/ml/datasets.html

Zbiory danych zostały ściągnięte i załadowane przy użyciu biblioteki *pandas* lub bezpośrednio załadowane za pomocą biblioteki *scikit-learn*.

Zbiory danych:

- Iris data set
- Wine data set
- Glass identification data set
- Pima diabetes data set

Ciekawostką jest, że w trakcie badania klasyfikatora i tworzenia sprawozdania ostatni ze zbiorów *Pima diabetes* został usunięty ze strony UCI ze przez ograniczenie uprawnień do udostępniania danego zbioru.


I'm sorry, the dataset "pima indians diabetes" does not appear to exist.

A note from the donor regarding Pima Indians Diabetes data:

"Thank you for your interest in the Pima Indians Diabetes dataset. The dataset is no longer available due to permission restrictions."

Rysunek 1 Wiadomość ze strony https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes mówiąca o braku dalszego dostępu do danego zbioru.

Poniżej zaprezentowano opis zbiorów. Opis ten pomoże w zrozumieniu danych, które będą analizowane. Dobre zrozumienie danych z którymi się pracuje jest niezbędną częścią do poprawnego przeprowadzenia badań.

Zbiór Iris

Jest to prawdopodobnie jeden z najbardziej znanych i podstawowych zbiorów danych przy problemach klasyfikacji i rozpoznawania wzorców.

Zbiór składa się ze 150 instancji, podzielonych na 3 równe zbiory po 50 klas każda.

Definicje atrybutów:

- Sepal – zielony płatek u dołu kielicha służący do ochrony kwiatu w trakcie kwitnięcia,
- Petal – płatek kwiatu, służący do przyciągania uwagi ptaków i insektów

Cechy zbioru zawierają cztery informacje:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

Ostatnią, piątą kolumną jest klasa mówiąca o typie irysa. Możliwe są trzy klasy:

1. Iris Setosa
2. Iris Versicolour
3. Iris Virginica

Zbiór Wine

Zbiór ten został skonstruowany w wyniku analizy składu chemicznego win stworzonych w tym samym rejonie Włoch lecz przy użyciu trzech różnych odmian uprawnych.

Zbiór składa się ze 178 instancji.

Definicje atrybutów oraz cechy zbioru:

1. Alcohol – alkohol
2. Malic acid – kwas jabłkowy
3. Ash – popiół
4. Alkalinity of ash – alkaliczność popiołu
5. Magnesium – magnez
6. Total phenols – całkowita zawartość fenoli
7. Flavonoids – flawonoidy
8. Nonflavanoid phenols – fenole nieflawonowe
9. Proanthocyanidins – proantocyjanidyny
10. Color intensity, intensywność koloru
11. Hue – odcień
12. OD280/OD315 of diluted wines - OD280 / OD315 rozcieńczonych win
13. Proline – Proline

Pierwszy atrybut w pliku zawierającym dane jest identyfikatorem klasy od 1 do 3.

Rozłożenie instancji klas jest następujące:

- Klasa 1 – 59 instancji,
- Klasa 2 – 71 instancji,
- Klasa 3 – 48 instancji.

Zbiór Glass identification

Zbiór powstał poprzez analizę składu chemicznego badanego szkła aby określić typ powstałego szkła oraz jego przeznaczenie.

Zbiór składa się z 214 instancji podzielonych na 6 klas.

Rozłożenie instancji klas jest następujące:

- Klasa 1 – 70 instancji,
- Klasa 2 – 76 instancji,
- Klasa 3 – 17 instancji,
- Klasa 4 - 13,
- Klasa 5 - 9,
- Klasa 6 - 29.

Definicje atrybutów oraz cechy zbioru:

1. Id – numer porządkowy
2. Refractive index – współczynnik załamania światła
3. Sodium – sód
4. Magnesium – magnez
5. Aluminium – glin
6. Silicon – krzem
7. Potassium – potas
8. Calcium – wapń
9. Barium – bar
10. Iron – żelazo

Zbiór Seeds

Zbiór reprezentuje atrybuty 3 różnych typów zbóż.

Zbiór składa się z 210 instancji podzielonych na 3 klasy.

Rozłożenie instancji klas jest następujące:

- Klasa 1 (Kama) – 70 instancji,
- Klasa 2 (Rosa) – 70 instancji,
- Klasa 3 (Canadian) – 70 instancji,

Definicje atrybutów oraz cechy zbioru. Wszystkie atrybuty są miarami nasion zboża:

1. Area – pole
2. Perimeter – obwód
3. Compactness – ścisłość
4. Length of kernel – długość nasiona
5. Width of kernel – szerokość nasiona
6. Asymmetry coefficient – współczynnik asymetrii
7. Length of kernel groove - długość rowka nasiona

Zbiór Pima diabetes

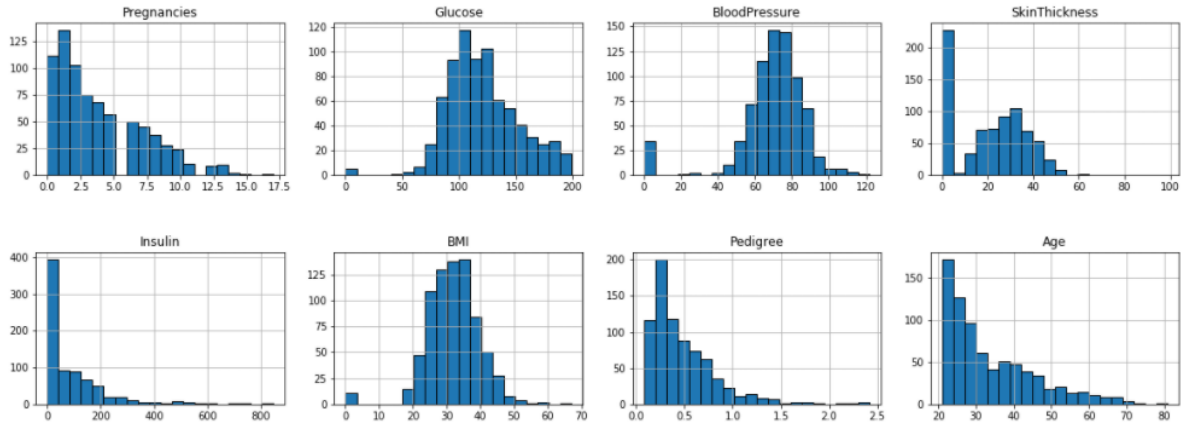
Celem zbioru jest umożliwienie zdiagnozowania czy dany pacjent ma cukrzycę, bazując na diagnostykach zamieszczonych w cechach zbioru. Wszyscy pacjenci

przebadani byli kobietami mającymi przynajmniej 21 lat oraz byli pochodzenia indiańskiego plemienia Pima.

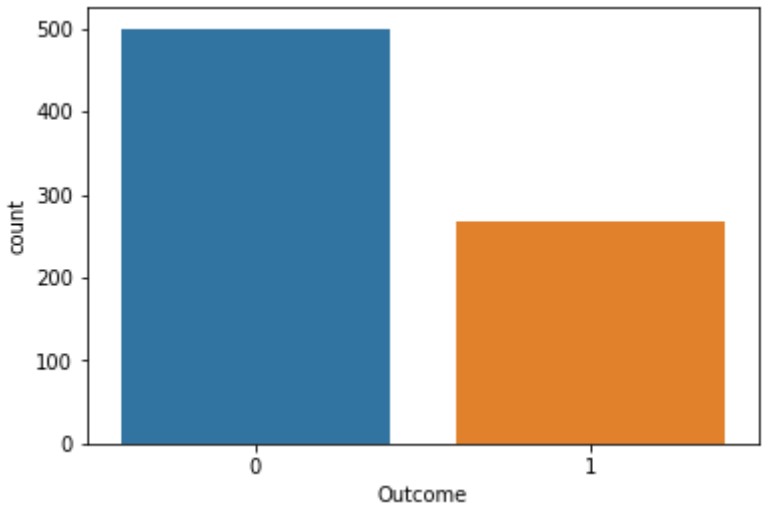
Zbiór składa się z 768 instancji posiadających dwie możliwe klasy 1 – oznaczające że zbadana osoba jest chora na cukrzycę, 0 – oznaczające że dana osoba nie jest chora na cukrzycę.

Definicje atrybutów oraz cechy zbioru:

1. Pregnancies – liczba ciąży
2. Glucose – poziom glukozy
3. Blood ressure – ciśnienie krwi
4. Skin thickness – grubość skóry
5. Insulin – poziom insuliny
6. BMI – body mass index
7. Diabetes pedigree function – funkcja pedigree
8. Age – wiek



Rysunek 2 Histogramy atrybutów danych zbioru Pima diabetes



Rysunek 3 Rozkład klas

3. Implementacja

Do implementacji użyty został skryptowy język programowania R oraz następujące biblioteki: Rweka, caret.

Tworzenie modelu:

Do stworzenia modelu k-means została użyta standardowa biblioteka R, natomiast to k-medoids i jego implementacji PAM została użyta biblioteka *cluster*.

Parametry wykorzystane w k-means i k-medoids

K (number of clusters) – Podstawowy parametr, który odpowiada za liczbę klastrów do stworzenia.

Nstart – liczba początkowych konfiguracji rozłożenia centroidów. Wybierany jest najlepszy i używany do działania algorytmu.

Parametry analizy k-means i k-medoids

Dunn index – Przyjmuje wartości od 0 do nieskończoności. Jest obliczany jako stosunek najmniejszej odległości między obserwacjami w różnych klastrach do największej odległości wewnątrz jednego klastra. Miara ta ma na celu zidentyfikowanie klastrów, które są zwarte (z niewielką rozbieżnością między członkami klastra) i dobrze rozdzielone (środki różnych klastrów są od siebie wystarczająco daleko oddalone) w porównaniu do wariancji wewnątrz klastra. **Im wyższy indeks Dunn'a tym lepsze są stworzone klastry.** Jednym z problemów z tą miarą jest koszt obliczeniowy, który rośnie wraz ze wzrostem liczby klastrów i wymiarów danych.

Silhouette coefficient – Przyjmuje wartości od -1 do +1. **Wysoka wartość mówi o tym że obiekt jest dobrze dopasowany do własnego klastra i źle do innych.** Mała wartość oznacza, że mamy za wiele lub zbyt mało klastrów. Współczynnik ten kontrastuje średnią odległość od elementów w tym samym klastrze ze średnią odległością od elementów w innych klastrach. Obiekty o wysokiej wartości Silhouette są uważane za dobrze zgrupowane, obiekty o niskiej wartości mogą być wartościami odstającymi. Indeks ten służy także do określania optymalnej liczby klastrów.

Davies-Bouldin index – Przyjmuje wartości od 0 do nieskończoności. *n* oznacza liczbę klastrów, *c* oznacza centroid klastra, *sigma* to średnia odległość wszystkich elementów w klastrze do jego centroida i *d* oznacza odległość pomiędzy dwoma centroidami. Algorytm, który wyprodukuje klastry z małą średnią odległością wewnątrz klastra i wysoką odległością między klastrami będzie miał niską wartość indeksu. **Im mniejsza wartość tej miary tym lepszy podział klastrów.**

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

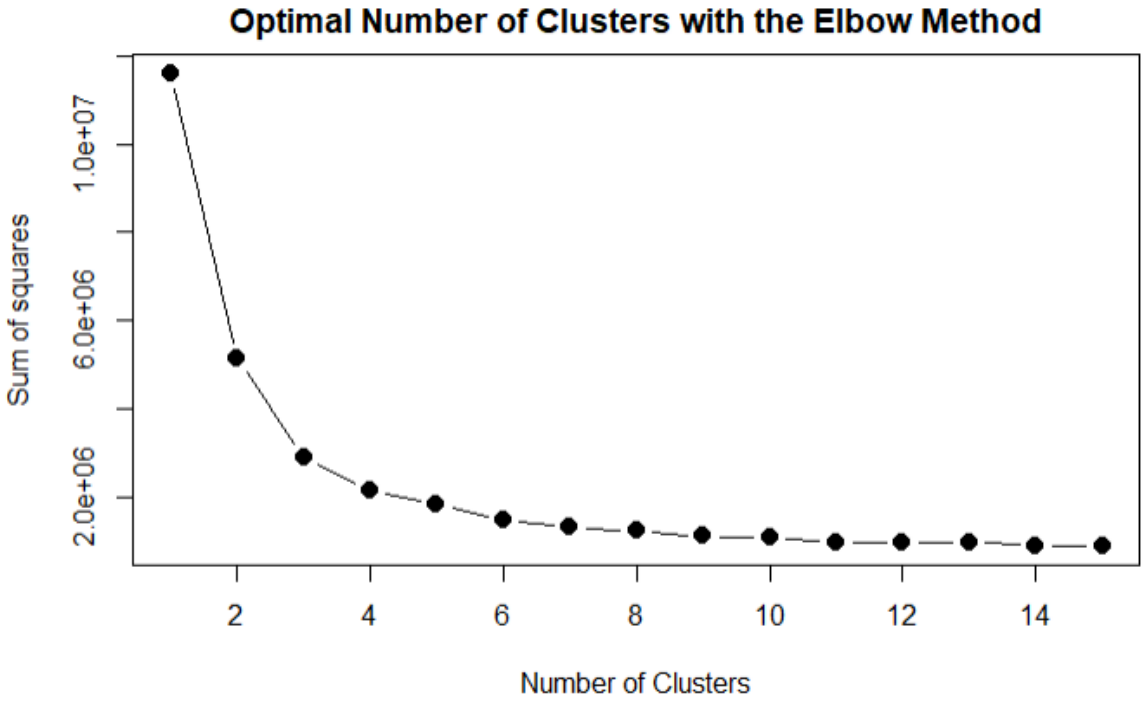
Rysunek 4 Wzór na davies-bouldin index

Dobór optymalnego parametru k – do wyboru optymalnego parametru liczby klastrów została użyta metoda łokcia “elbow method”. Mówi ona o o tym aby wypróbować kilka różnych k i poszukać miejsca na grafie, w którym przestajemy uzyskiwać duże zyski (w tym przypadku zmniejszamy sumę odległości) informacji i większa liczba klastrów przestaje dawać duże zmiany. To miejsce najczęściej jest zakrzywieniem w grafie, które przypomina łokieć.

4. Analiza wyników zbiorów danych

a. Analiza wyników zbioru Pima Diabetes

Elbow-method



Rysunek 5 Wybór optymalnego parametru metodą elbow dla zbioru Pima diabetes

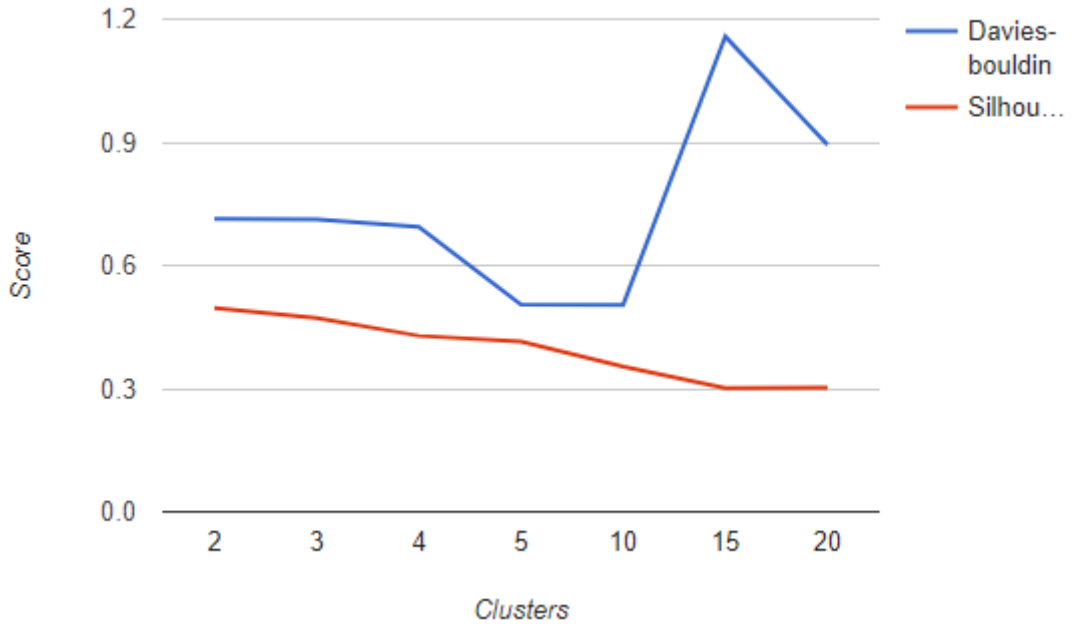
W przypadku zbioru pima diabetes optymalna wartość parametru k według metody łokcia znajduje się w granicy 3 klastrów. Późniejsze zwiększanie liczby klastrów nie daje aż tak dużego zysku.

K-means

Wartość Nstart została ustawiona na 100.

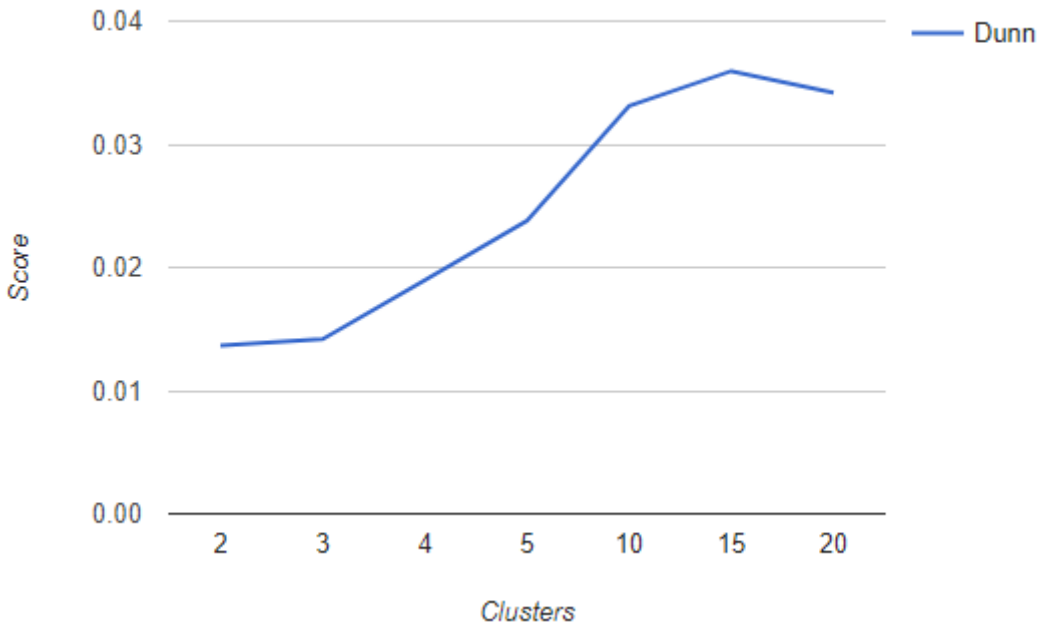
K	Dunn	Silhouette	Davies-bouldin

2	0.0136	0.496	0.713
3	0.0141	0.471	0.712
4	0.0189	0.428	0.693
5	0.0237	0.414	0.504
10	0.0331	0.353	0.503
15	0.0359	0.301	1.157
20	0.0341	0.301	0.893



Rysunek 6 Wyniki Davies-bouldin i silhouette dla pima diabetes przy użyciu k-means

W odróżnieniu od wyniku indeksu Dunn'a i Silhouette z miary Davies-bouldin wynika, że w przypadku 15 klastrów mamy najgorsze wyniki.

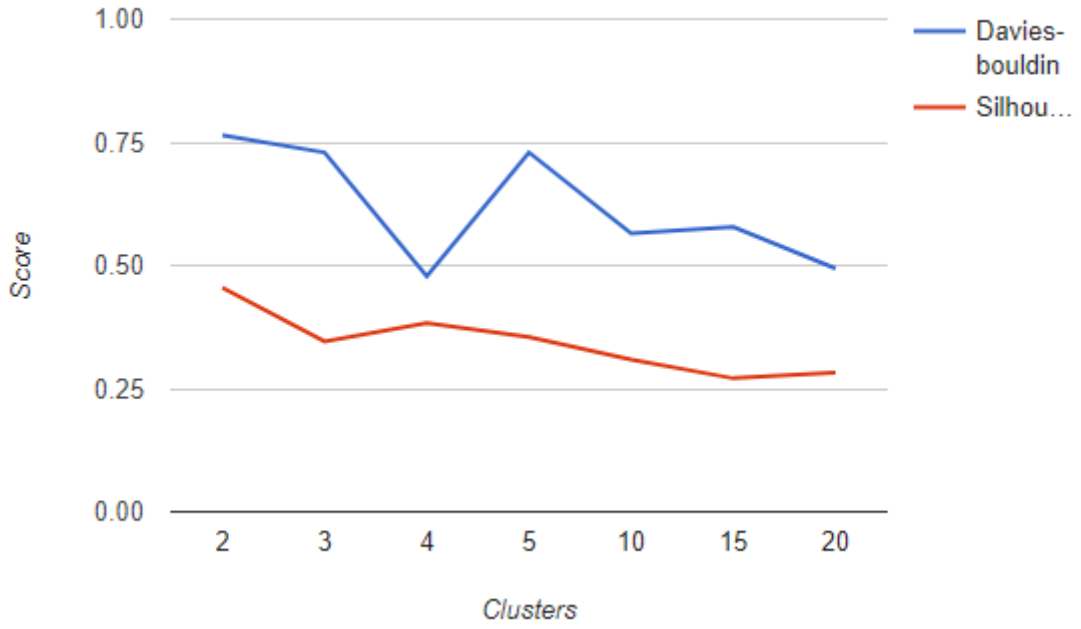


Rysunek 7 Wyniki Dunn dla pima diabetes przy użyciu k-means

W brzypadku zbioru Pima diabetes widać, że najlepsze rozłożenie klastrów dla indeksu Dunna jest przy 15 klastrach, a najmniejsze przy bardzo małej liczbie klastrów.

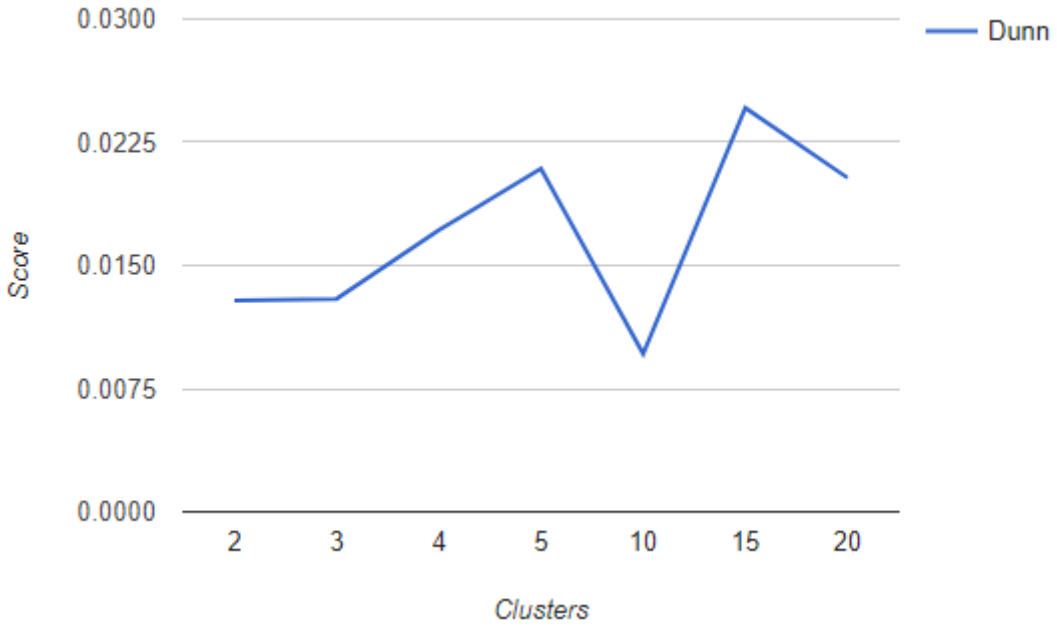
K-medoids

K	Dunn	Silhouette	Davies-bouldin
2	0.0128	0.454	0.764
3	0.0129	0.345	0.729
4	0.0171	0.382	0.478
5	0.0208	0.354	0.729
10	0.009	0.309	0.564
15	0.024	0.270	0.577
20	0.020	0.282	0.494



Rysunek 8 Wyniki Davies-bouldin i Slhouette dla pima diabetes przy użyciu PAM

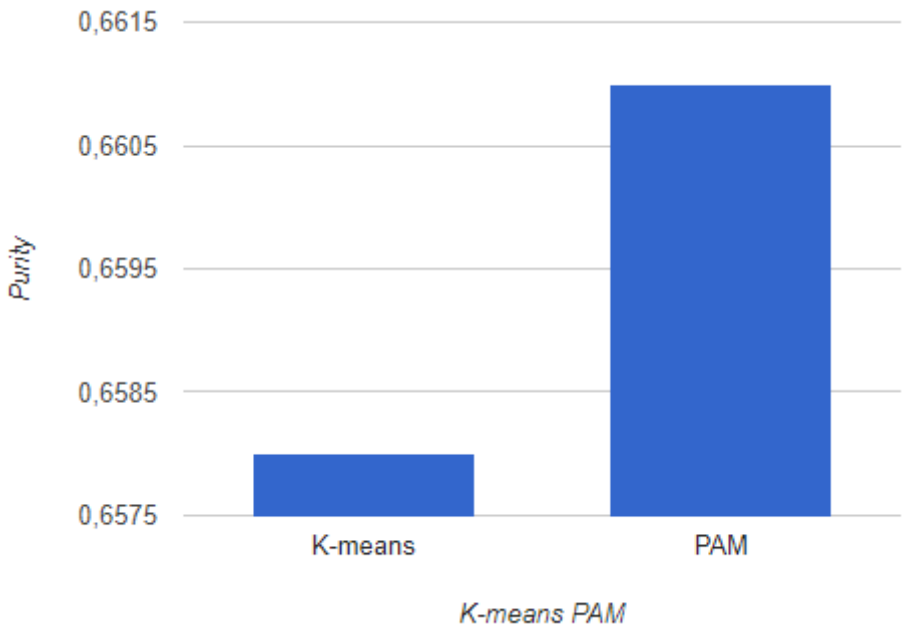
W przypadku mierzenia wyników za pomocą algorytmu PAM miary są do siebie bardziej zbliżone i nie występują aż tak wysokie skoki parametru davies-bouldin jak w przypadku k-means.



Rysunek 9 Wyniki Dunn dla pima diabetes przy użyciu PAM

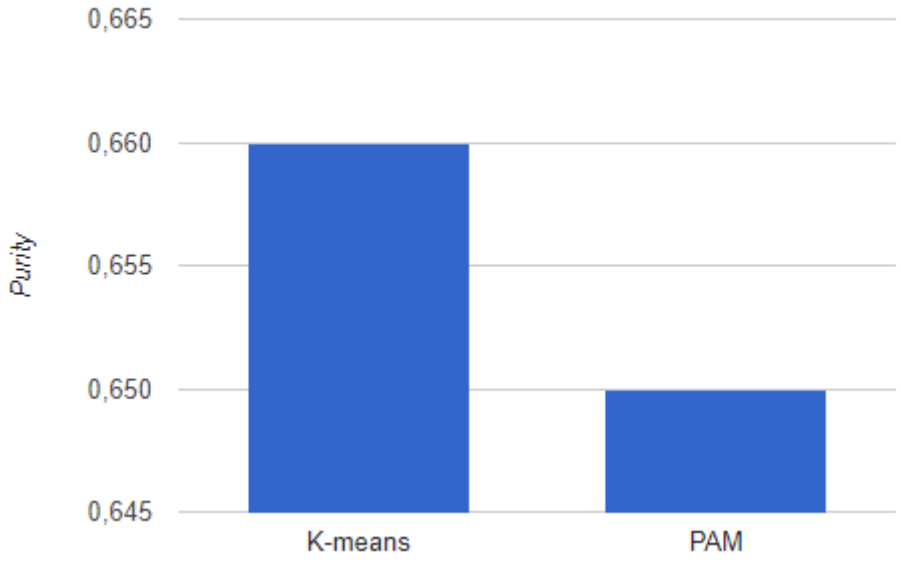
Podobnie jak w przypadku algorytmu k-medoids najlepszy wynik uzyskano przy około 15 klastrach.

Purity bez normalizacji



Rysunek 10 Wykres miary Purity przy parametrze k=3 dla zbioru Pima diabetes

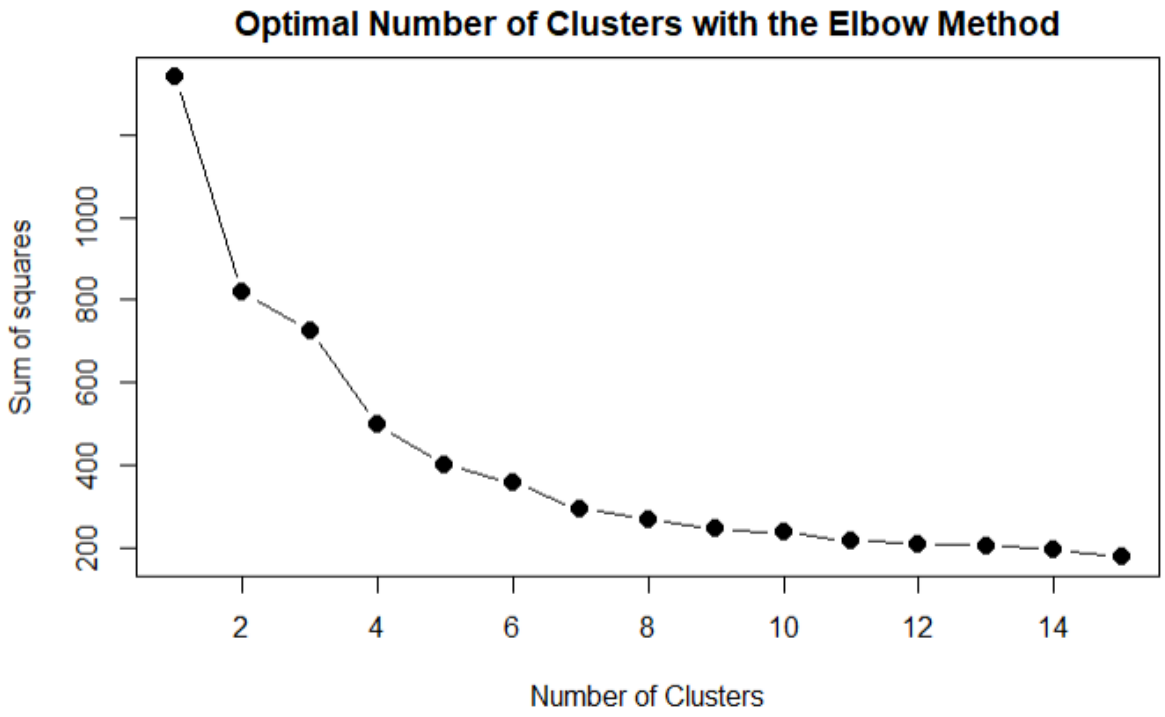
Purity z normalizacją



Pomimo tego liczba klastrów (3) była zbliżona do liczby faktycznych klas zbioru Pima diabetes (2) wynik miary Purity jest bardzo niski bo na poziomie 65-66%.

b. Analiza wyników zbioru Glass dataset

Elbow-method

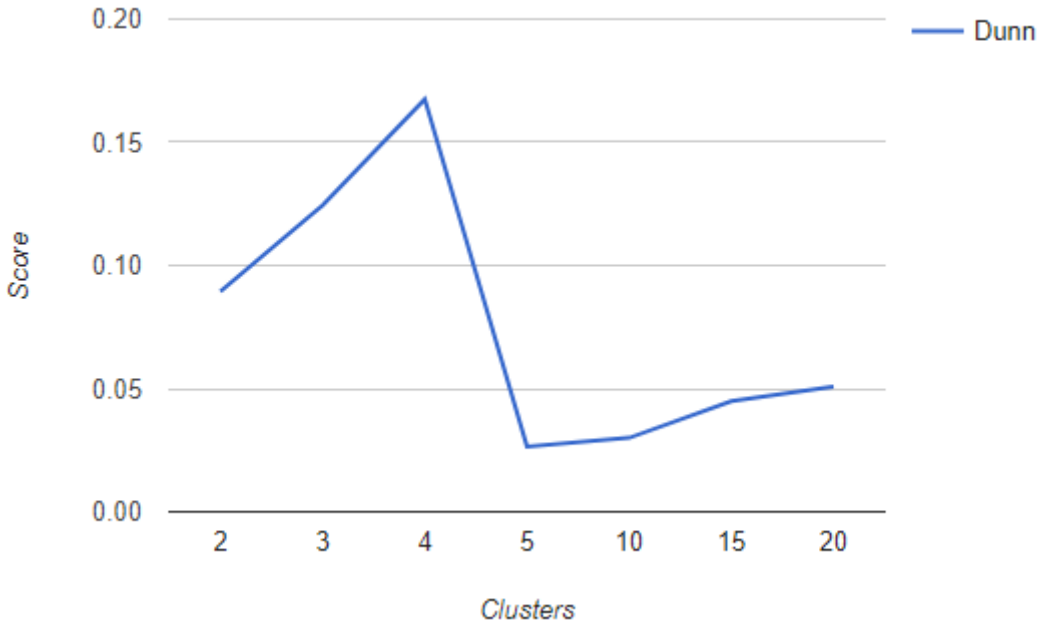


Rysunek 11 Wybór optymalnego parametru metodą elbow dla zbioru glass dataset

Jak widać nie w każdym przypadku metoda elbow działa odpowiednio dobrze. W przypadku tego zbioru nie ma dobrze widocznej części łokcia na grafie.

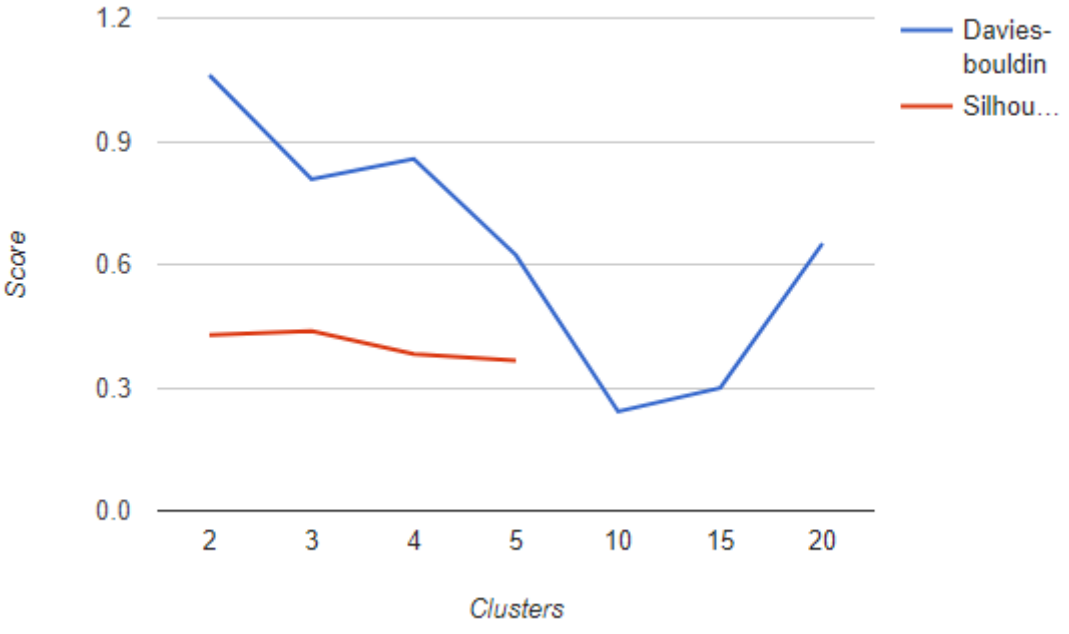
K-means

K	Dunn	Silhouette	Davies-bouldin
2	0.089	0.427	1.061
3	0.124	0.437	0.807
4	0.167	0.381	0.856
5	0.026	0.365	0.622
10	0.029	NaN	0.241
15	0.044	NaN	0.299
20	0.050	NaN	0.651



Rysunek 12 Wyniki Dunn dla glass przy użyciu k-means

Indeks Dunna dla zbioru glass wskazuje na to, że najlepszą liczbą klastrów są 4 klastry. Zgadza się to z przybliżonym wynikiem dla metody elbow.

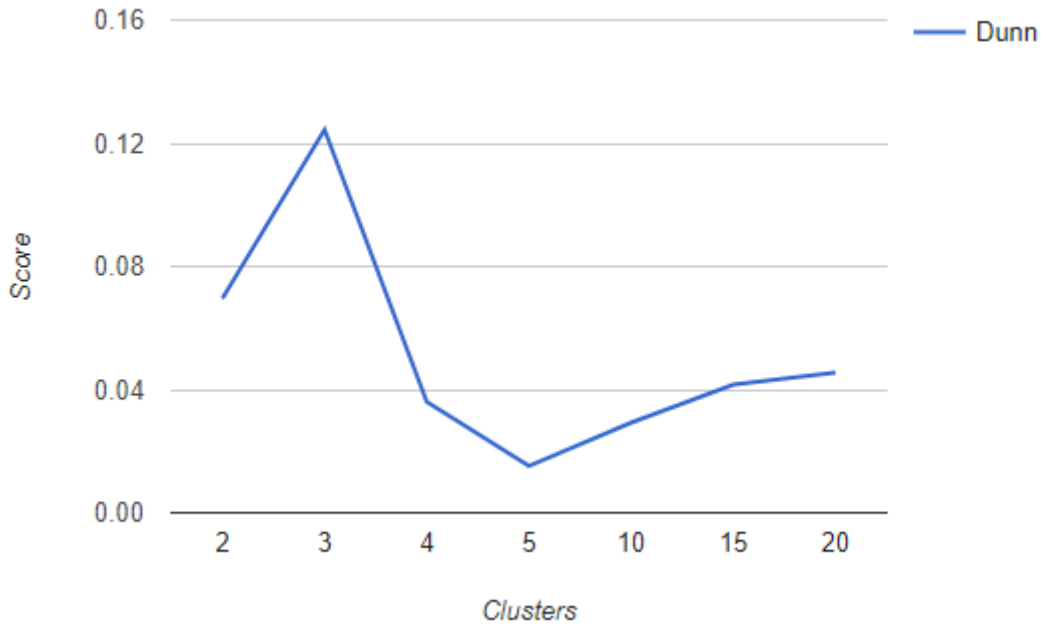


Rysunek 13 Wyniki Davies-bouldin i silhouette dla glass przy użyciu k-means

Dla zbioru glass widać, że nie możliwe było obliczenie wartości parametru Silhouette dla wartości 10 klastrów i powyżej nie było możliwe. Jednakże najlepsza wartość parametru davies-boulding wskazywała na liczbę 10 klastrów.

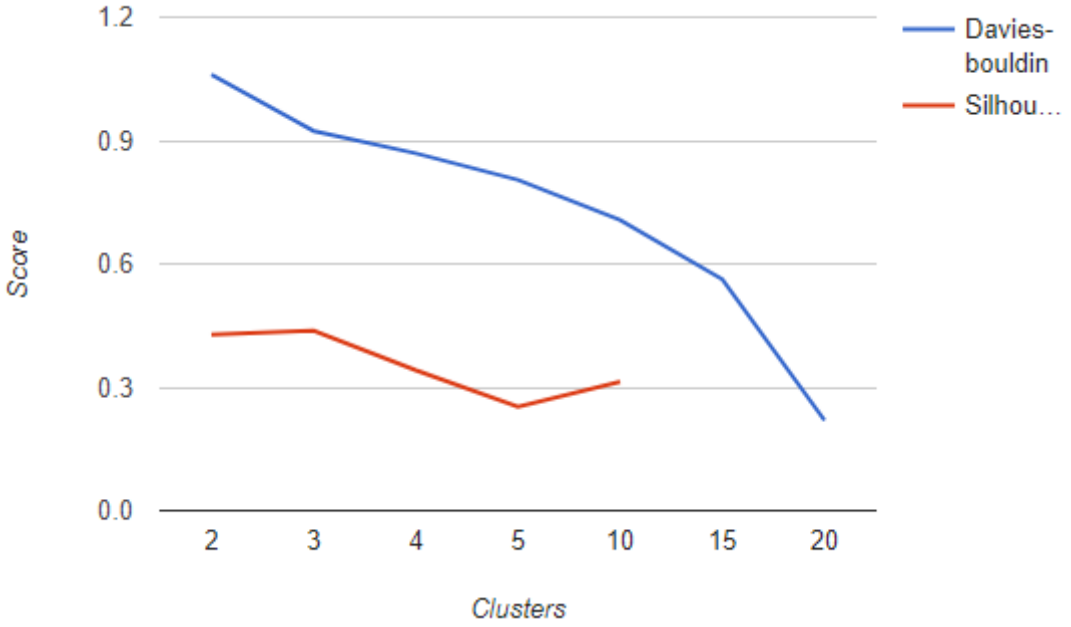
K-medoids

K	Dunn	Silhouette	Davies-bouldin
2	0.069	0.428	1.060
3	0.124	0.437	0.923
4	0.036	0.341	0.868
5	0.015	0.253	0.804
10	0.029	0.312	0.706
15	0.041	NaN	0.562
20	0.045	NaN	0.219



Rysunek 14 Wyniki Dunn dla glass przy użyciu PAM

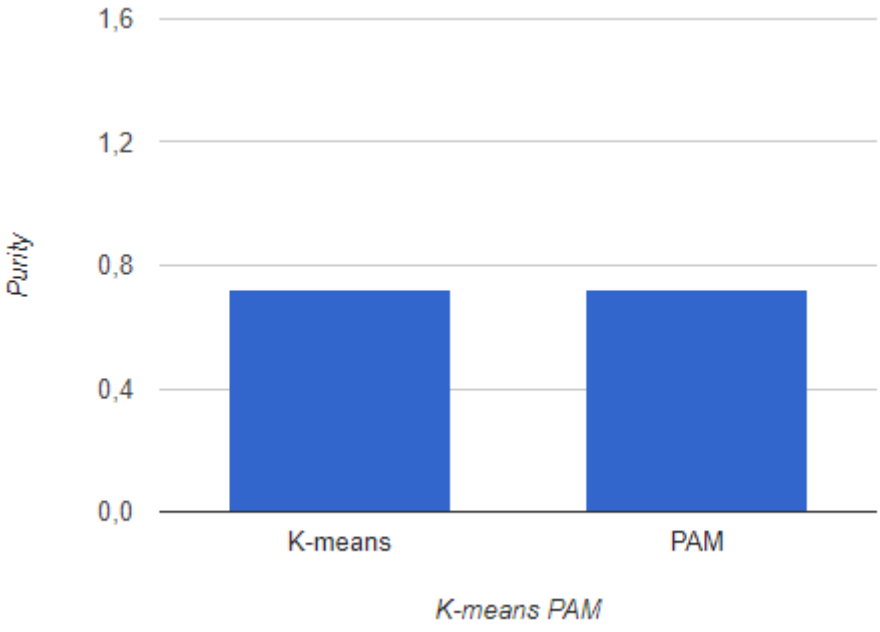
Tak jak w przypadku k-means indeks Dunna najwyższy jest przy około 3 klastrach, zatem blisko wyniku 4 dla k-means.



Rysunek 15 Wyniki Davies-bouldin i silhouette dla glass przy użyciu PAM

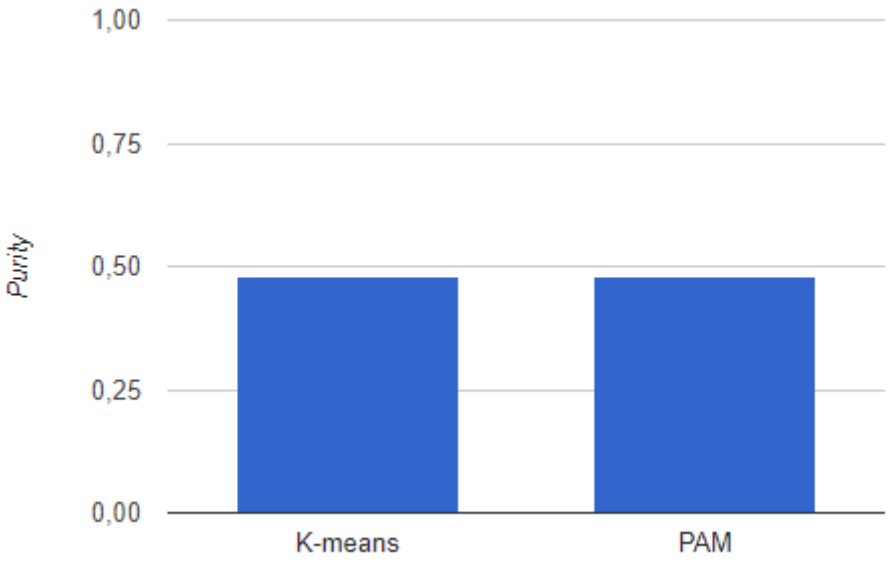
Tak jak w przypadku k-means nie można było obliczyć indeksu Silhouette dla wszystkich badanych miar liczby klastrów, jednakże można było policzyć ją jeszcze dla k=10.

Purity bez normalizacji



Rysunek 16 Wykres miary Purity przy parametrze k=4 dla zbioru Glass

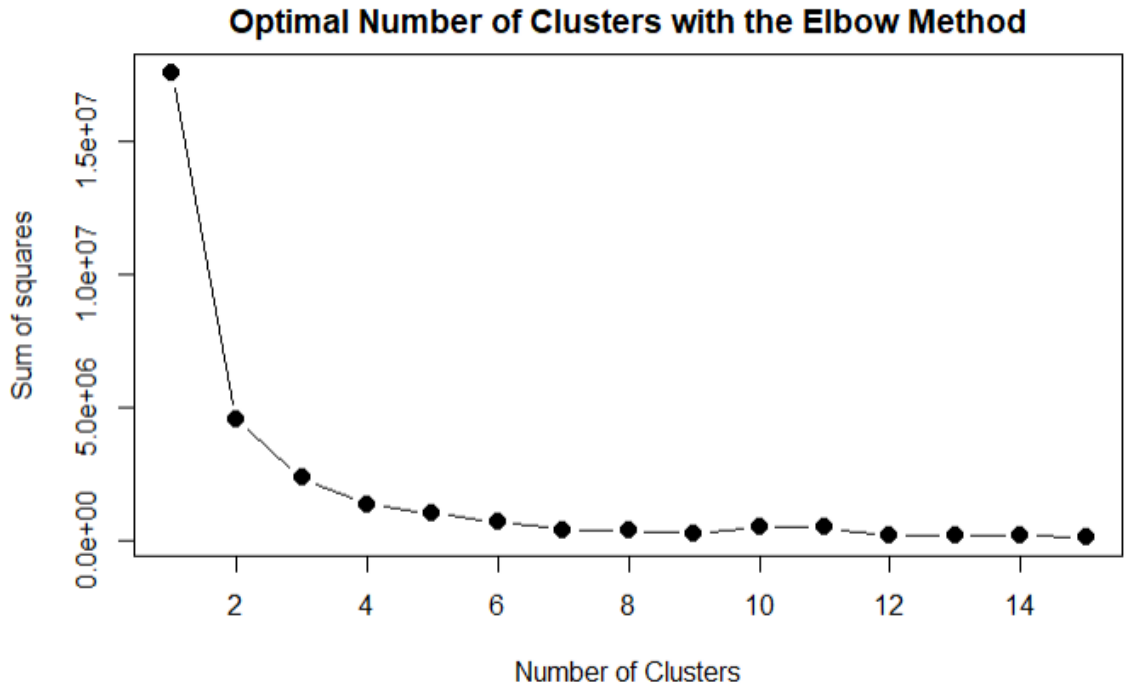
Purity z normalizacją



W przypadku zbioru Glass uzyskano identyczne wyniki dla algorytmu k-means oraz PAM. Wyniki uzyskano na poziomie 78%.

c. Analiza wyników zbioru Wine dataset

Elbow-method

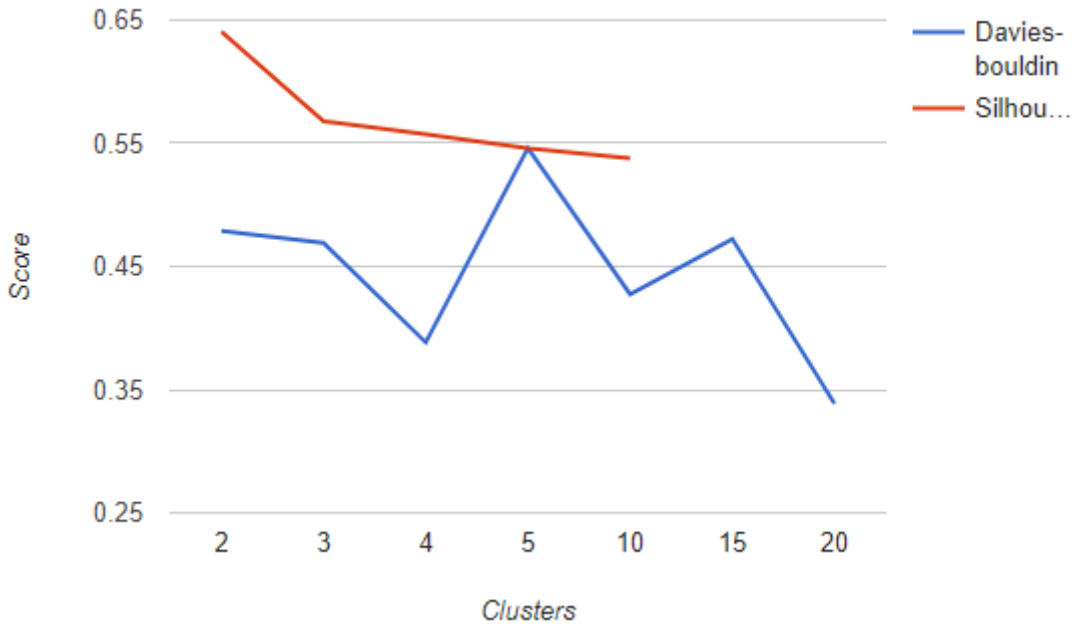


Rysunek 17 Wybór optymalnego parametru metodą elbow dla zbioru Wine dataset

Optymalna wartość liczby klastrów dla zbioru Wine oszacowana jest na około 3.

K-means

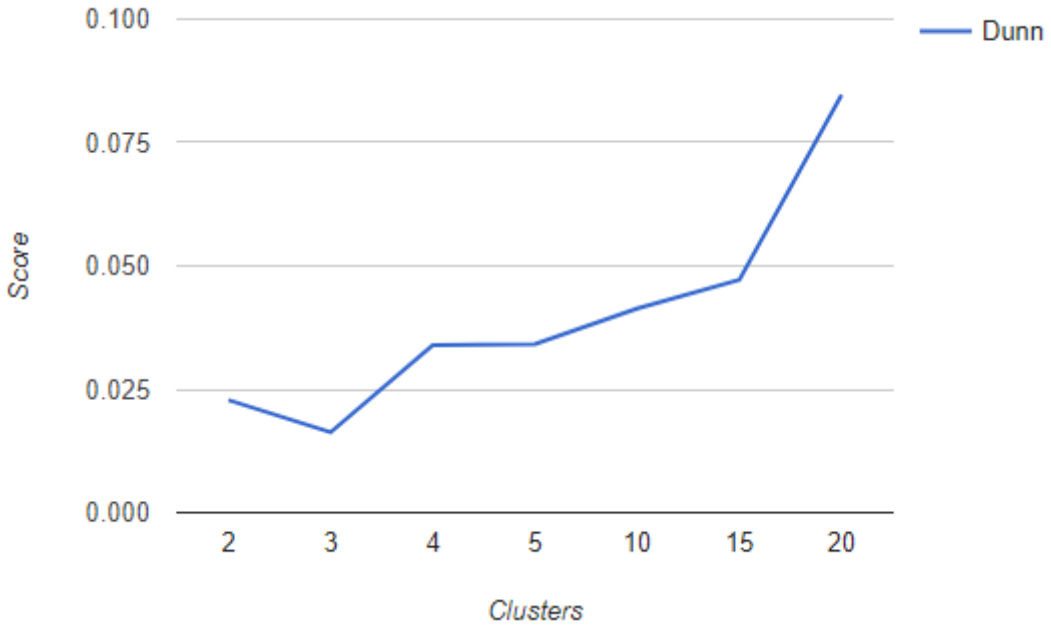
K	Dunn	Silhouette	Davies-bouldin
2	0.022	0.640	0.478
3	0.016	0.567	0.469
4	0.033	0.557	0.388
5	0.034	0.546	0.546
10	0.041	0.537	0.427
15	0.047	NaN	0.472
20	0.084	NaN	0.338



Rysunek 18 Wyniki Davies-bouldin i silhouette dla Wine dataset przy użyciu k-means

W przypadku algorytmu k-means przy badaniu zbioru Wine z parametru davies-boulding nie jesteśmy w stanie wiele odczytać ponieważ drastycznie zmieniają się

jego wartości przy każdym parametrze k. Jednakże tak jak obliczono z metody elbow największy zysk miary indeksu Silhouette został uzyskany do k=3.

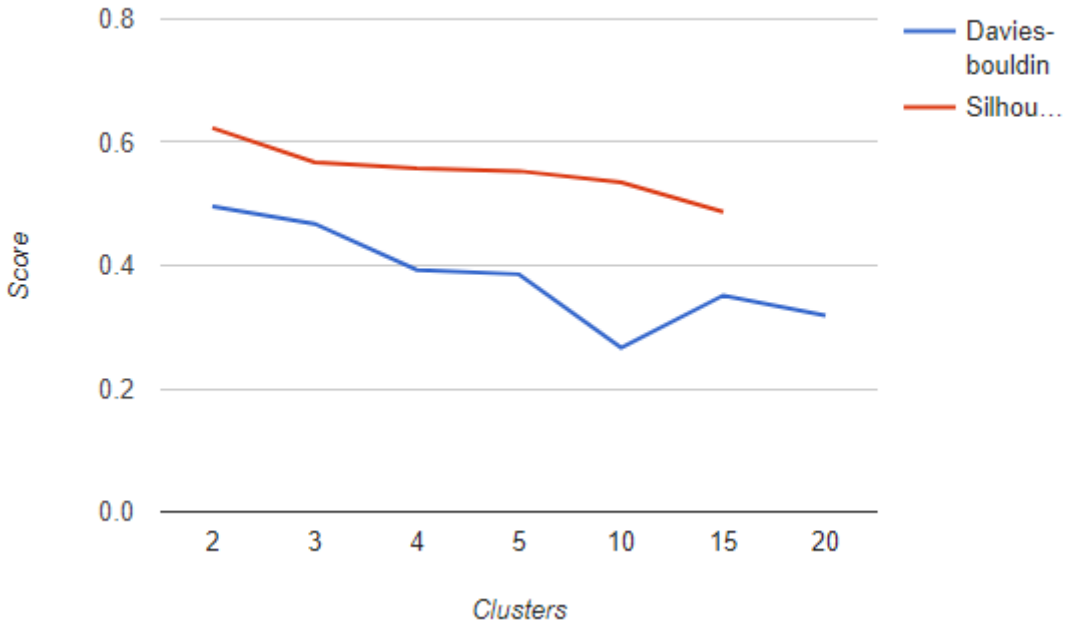


Rysunek 19 Wyniki Dunn dla Wine dataset przy użyciu k-means

W przypadku indeksu Dunn'a z każdym zwiększeniem parametru k, oprócz wartości 3 miara parametru rosła. Dziwnym jest, że właśnie w przypadku gdzie dla innych parametró k=3 było dużym zyskiem, tutaj jest ono jedynym spadkiem wartości miary Dunn'a.

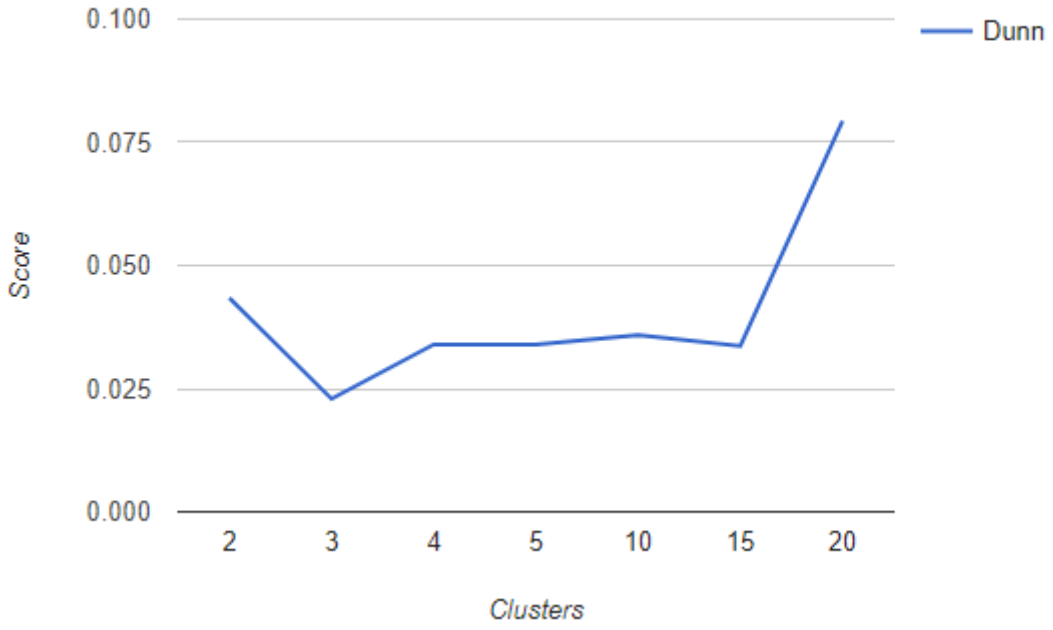
K-medoids

K	Dunn	Silhouette	Davies-bouldin
2	0.043	0.623	0.496
3	0.022	0.567	0.467
4	0.033	0.557	0.392
5	0.033	0.553	0.385
10	0.035	0.535	0.266
15	0.033	0.487	0.351
20	0.079	NaN	0.318



Rysunek 20 Wyniki Davies-bouldin i silhouette dla Wine dataset przy użyciu PAM

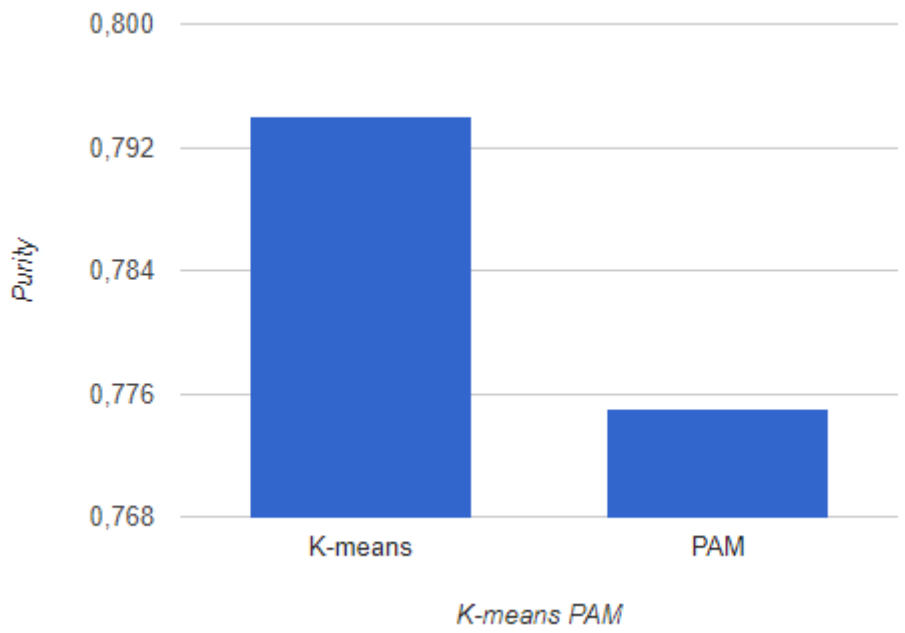
Dla zbioru Wine i użytego algorytmu PAM widzimy zdecydowanie bardziej stabilną zmianę miary davies-bouldin niż w przypadku algorytmu k-means.



Rysunek 21 Wyniki Dunn dla Wine dataset przy użyciu PAM

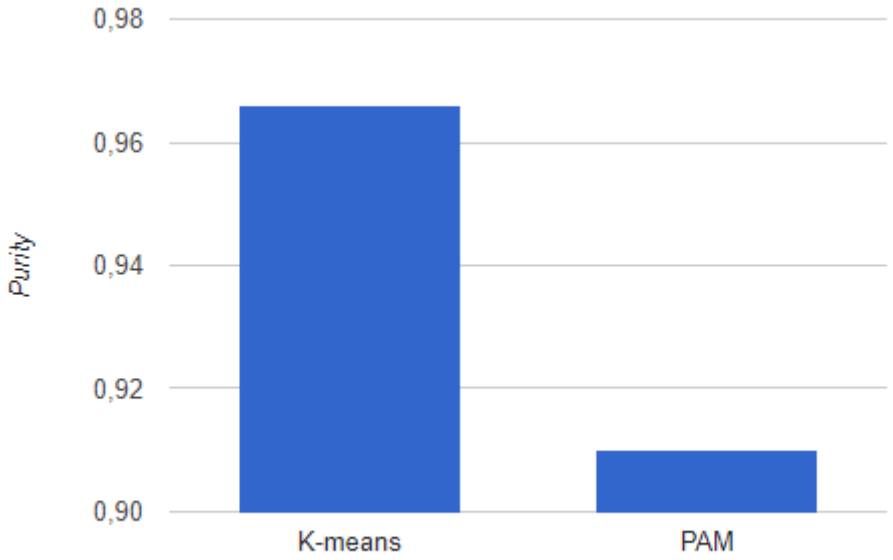
Przy algorytmie k-medoids jak i k-means indeks Dunn'a jest najmniejszy, czyli najgorszy, w przypadku teoretycznie optymalnego indeksu k, a zwiększe się bardzo szybko w przypadku dużej miary k.

Purity bez normalizacji



Rysunek 22 Wykres miary Purity przy parametrze k=3 dla zbioru Wine

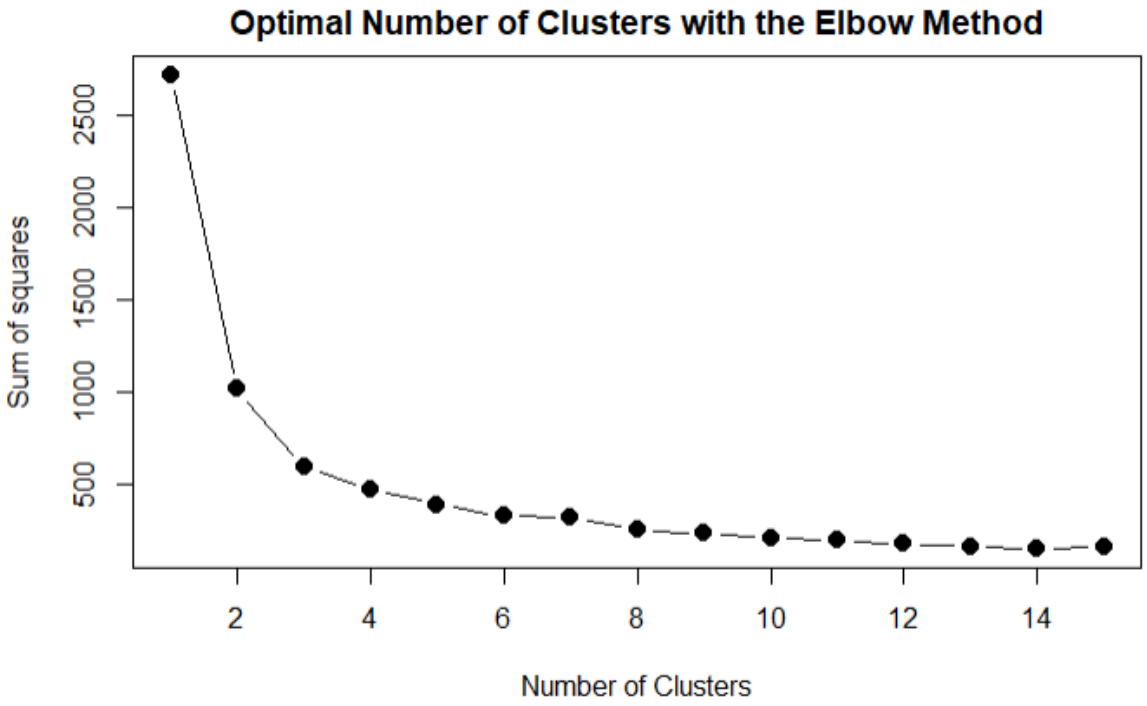
Purity z normalizacją



Dla zbioru Wine uzyskano wyniki na poziomie 77-79%. Wynik dla k-means był wyższy niż dla algorytmu PAM lecz jedynie o około 2%.

d. Analiza wyników zbioru Seeds dataset

Elbow-method

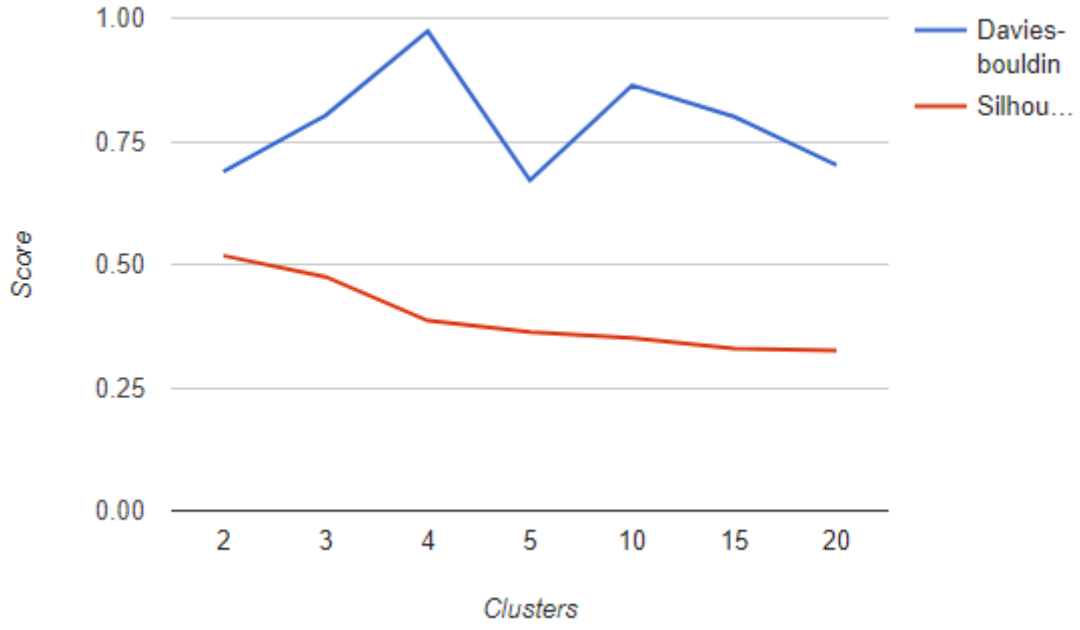


Rysunek 23 Wybór optymalnego parametru metodą elbow dla zbioru Seeds dataset

Z metody elbow wynika, że optymalny parametr k dla zbioru Seeds znajduje się przy liczbie klastrów równej 3. Późniejsze zwiększanie liczby klastrów nie przynosi dużego zysku.

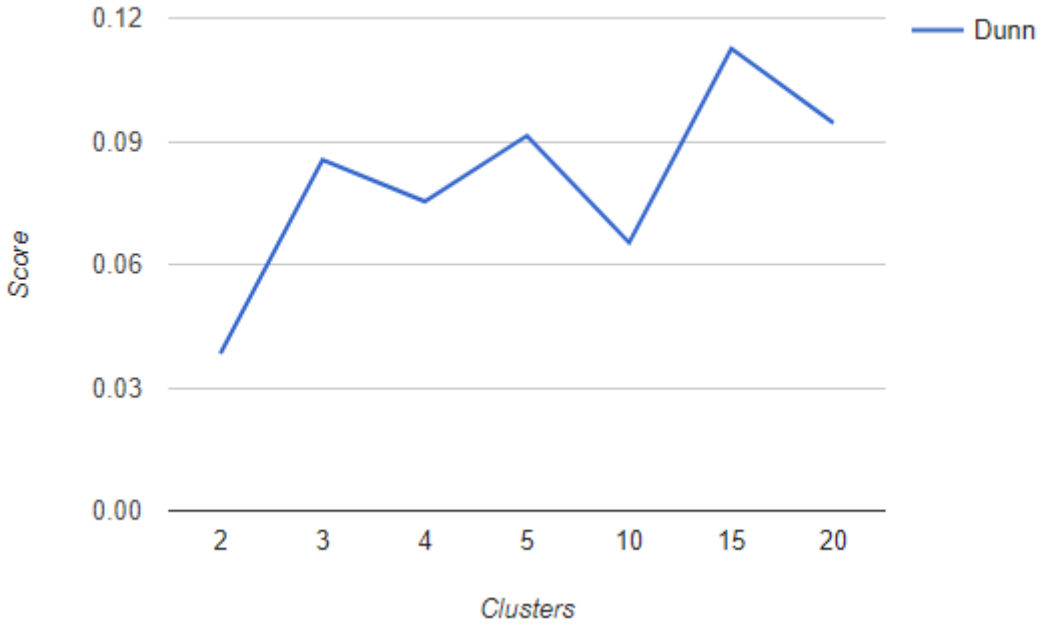
K-means

K	Dunn	Silhouette	Davies-bouldin
2	0.038	0.517	0.688
3	0.085	0.474	0.802
4	0.075	0.385	0.973
5	0.091	0.362	0.670
10	0.065	0.350	0.863
15	0.112	0.329	0.799
20	0.094	0.324	0.701



Rysunek 24 Wyniki Davies-bouldin i silhouette dla Seeds dataset przy użyciu k-means

Najlepsze wyniki dla miary Silhouette uzyskano w przypadku małej liczby klastrów (2,3), późniejsze zwiększanie liczby klastrów powodowało spadek wyników tej miary. Dla davies-bouldin największą zmianę miary widzimy dla k=5, który jest najlepszym wynikiem dla zbioru Seeds.

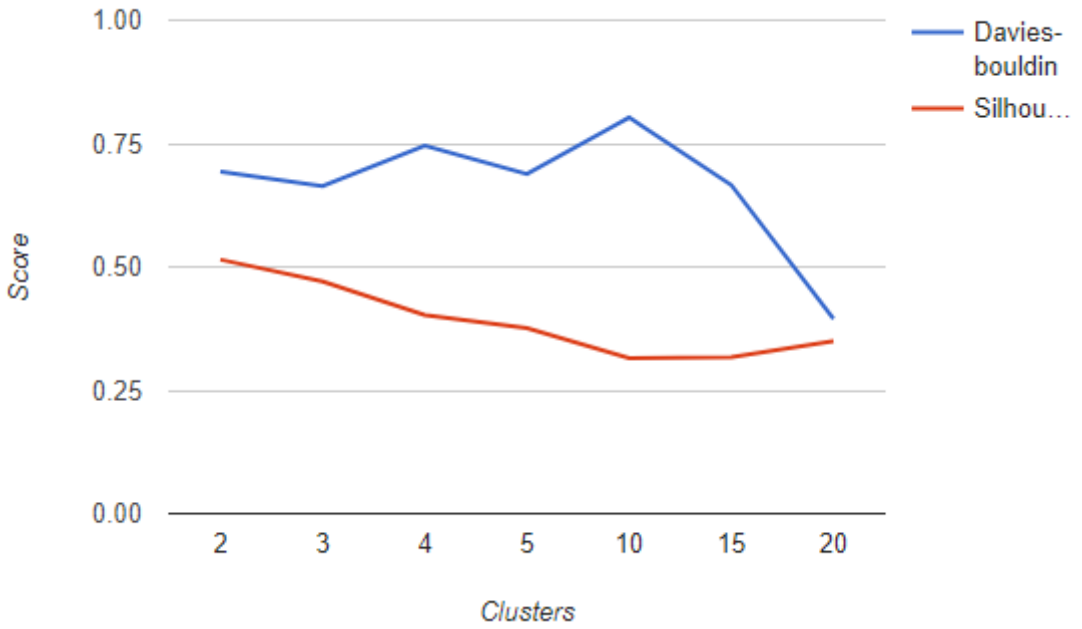


Rysunek 25 Wyniki Dunn dla Seeds dataset przy użyciu k-means

Przy obliczaniu miary Dunn'a widać duże zmiany indeksu przy zmienie parametru k. Jednakże tak jak pokazuje metoda elbow największy zysk uzyskujemy dla wartości k=3, późniejsze zmiany dają niewielki spadek lub zwiększenie parametru.

K-medoids

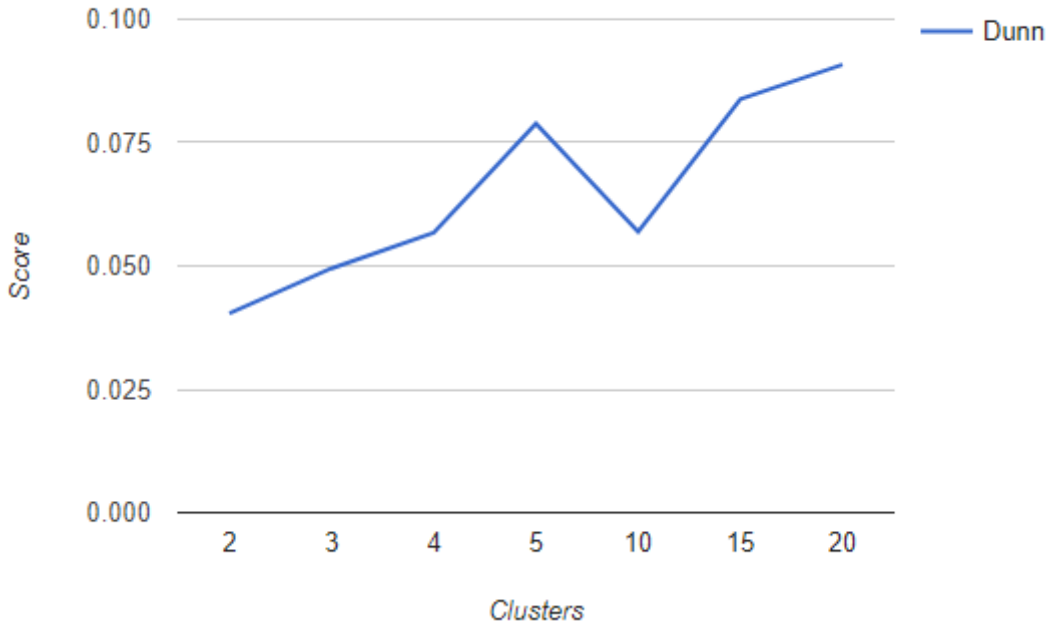
K	Dunn	Silhouette	Davies-bouldin
2	0.0403	0.514	0.693
3	0.0495	0.470	0.664
4	0.056	0.402	0.746
5	0.078	0.376	0.688
10	0.056	0.315	0.802
15	0.083	0.317	0.666
20	0.091	0.349	0.394



Rysunek 26 Wyniki Davies-bouldin i silhouette dla Seeds dataset przy użyciu PAM

Wyniki indeksu Davies-bouldin dla zbioru Seeds znajduje się przy dużej liczbie klastrów, gdy zwiększony został limit klastrów do 15 i 20 uzyskano duży spadek

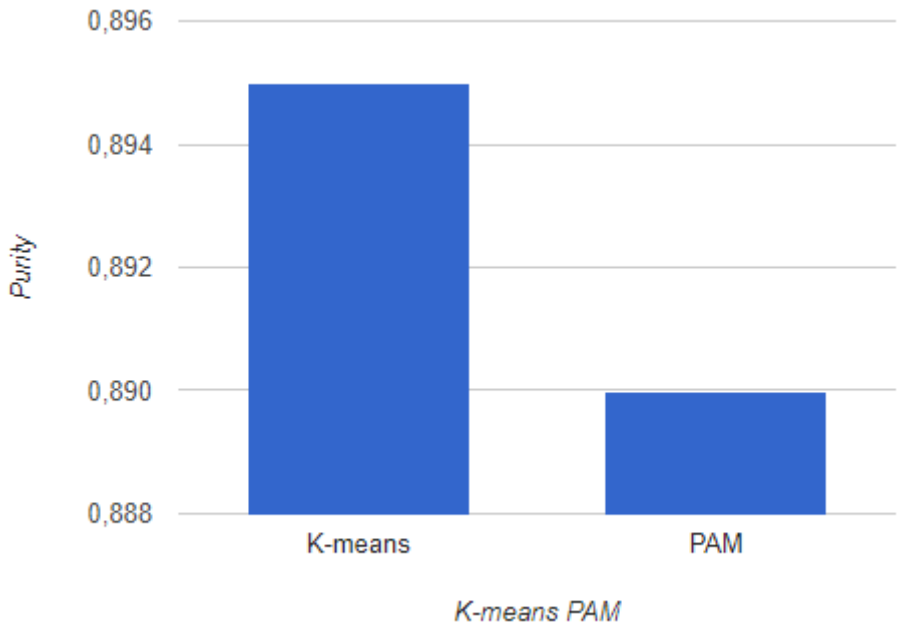
miary co przekłada się na jego lepszy wynik. Tak jak w poprzednich analizach, najlepsze wyniki Silhouette wypadają przy niskiej liczbie klastrów.



Rysunek 27 Wyniki Dunn dla Seeds dataset przy użyciu PAM

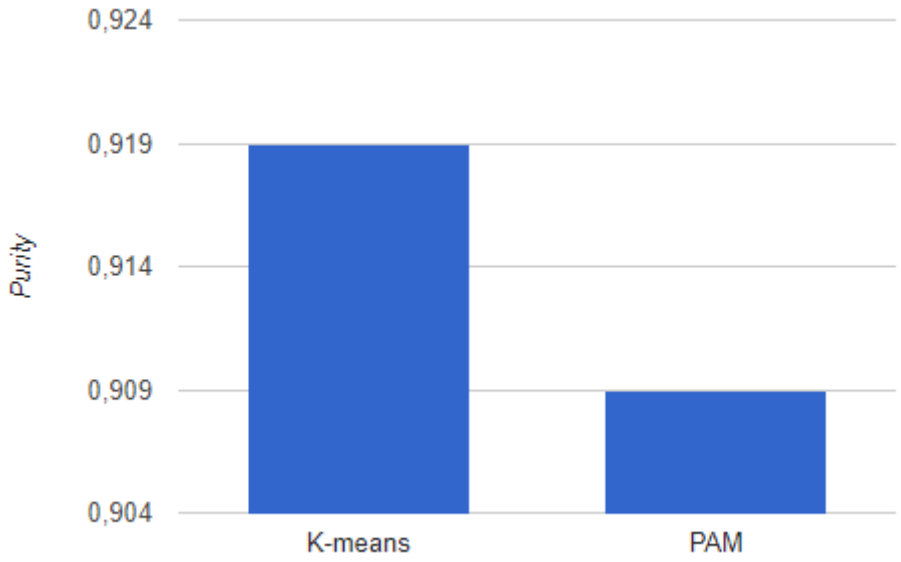
W przypadku metody PAM widać dosyć regularny wzrost wyniku miary Dunn'a, z wyłączeniem przypadku dla 10 klastrów.

Purity bez normalizacji



Rysunek 28 Wykres miary Purity przy parametrze k=3 dla zbioru Seeds

Purity z normalizacją



Optymalna liczba klastrów w przypadku zbioru Seeds ustalona została za pomocą metody Elbow na k=3, co zgadza się z liczbą klas z zbiorze. Pomimo tego miara Purity nie była wyższa niż 89% w obu przypadkach.

5. Podsumowanie

W przypadku porównywania algorytmów k-medoids i jego implementacji Partition around medoids z algorytmem k-means bardziej płynną zmianę wartości badanych parametrów. Jest to spowodowane między innymi przez to, że obserwacje, które mocno odbiegają od innych (tzw. outliers), mają duży wpływ w przypadku algorymtu k-means. Jest to spowodowane tym, że miara którą posługuje się algorytm k-means w przypadku obliczania nowego centroida w klastrze jest odległość pomiędzy punktami.

Najwyższe wyniki miary Silhouette najczęściej osiągano w przypadku małej liczby klastrów. Najczęściej miara była stabilna do około 4 klastrów, potem wynik zaczynał spadać.

W większości analizowanych przypadków miary nie wskazywały na tą samą liczbę klastrów jako optymalną wartość. Zdarzało się także, że miary wskazywały na to, że dla takiego samego k jedna miara miała wynik najgorszy, a inna najlepszy.