POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: INFORMATYKA

Systemy uczące się
Laboratorium

Klasteryzacja. Metoda k-means oraz PAM.

AUTOR:
Bartosz Kardas

PROWADZĄCY:
dr inż. Paweł Myszkowski

OCENA PRACY:

WROCŁAW, 2017

Spis treści

Spis rysunków

Spis tabel

Rozdział 1

Opis algorytmów

1.1. K-means

Algorytm K-means to jeden z najprostszych algorytmów klasteryzacji. Polega on na interacyjnym dobieraniu centroidów dopóki nie spełniono warunku stopu. Poniżej znajduje się szczegółowy opis algorytmu:

1. Losowe wybranie k centroidów.
2. Dla każdego elementu ze zbioru wyznaczenie najbliższego centroidu i przypisanie go do klastra.
3. Aktualizacja centroidów na podstawie wszystkich wartości w klastrze (wyznaczenie obiektu o średnich wartościach wszystkich atrybutów wewnątrz klastra)
4. Powtarzanie kroków 1-3 zadaną ilość iteracji, bądź dopóki centroidy zmieniają swoje położenie.

Istnieją różne odmiany tego algorytmu:

- Lloyd-Forgy – iteracyjne wyznaczanie centroidów na podstawie najbliższych sąsiadów
- MacQueen – polega na iteracyjnym przesuwaniu centroidów w stronę średniej ze zbiorów Vornoi
- Hartigan-Wong – minimalizuje sumę kwadratów w grupach

1.2. PAM - Partitioning Around Medoids

W przypadku algorytmu PAM, klasteryzacja odbywa się za pomocą tzw. medoidów, czyli elementów ze zbioru, których zróżnicowanie względem wszystkich elementów w klastrze jest najmniejsze. Algorytm PAM podobnie jak k-means jest algorytmem iteracyjnym. Poniżej szczegółowo przedstawiono proces wyznaczania klastrów za pomocą tego algorytmu.

1. Wylosować k elementów ze zbioru danych i oznaczyć je jako obecne medoidy.
2. Dla każdego elementu ze zbioru przypisać mu najbliższy medoid według danej metryki.
3. Dla każdego wyznaczonego klastra:
    1. Dla każdego elementu wewnątrz klastra:
        1. Oznaczyć element przez medoid i obliczyć sumaryczną odległość od pozostałych elementów w klastrze (funkcja kosztu).
    2. Wybrać nowy medoid spośród kandydatów z 3.1.1. wewnątrz klastra, którego funkcja kosztu jest najmniejsza.
4. Powtarzać kroki 2-3 dopóki medoidy zmieniają się.

W powyższym algorytmie można zastosować różne metryki. Najpopularniejszymi z nich są:

- metryka euklidesowa
- metryka miejska (metryka Manhattan)

W badaniach uwzględniono różne metryki w algorytmie PAM.

1.3. Miary klasteryzacji

1.3.1. Purity

Jeśli zbiór danych ma określone klasy, można wyznaczyć miarę dopasowania klastrów do obecnych klas wewnątrz zbioru. Do tego celu wykorzystuje się miarę purity. Poniżej przedstawiono wzór dzięki któremu można obliczyć tę miarę.

$$purity(\Omega, C) = \frac{1}{N} \sum_{k} max_j |\omega_k \cap c_j|$$

,gdzie ω_k określa k-ty klaster, a c_j określa j-tą klasę. Innymi słowy miara ta to sumaryczna ilość wystąpień pewnej klasy wewnątrz danego klastra podzielona przez ilość instancji w zbiorze.

1.3.2. Silhouette

Miara silhouette określa jak dobrze położony jest dany obiekt wewnątrz klastra. Do jej obliczenia należy wyznaczyć współczynniki $a(i)$ oraz $b(i)$, których znaczenie opisano poniżej:

- $a(i)$ - średnia odległość i-tego obiektu od pozostałych obiektów należących do tego samego klastra
- $b(i)$ - minimalna średnia odległość i-tego obiektu od pozostałych klastrów

Dla tak zdefiniowanych wartości $a(i)$ i $b(i)$ można obliczyć miarę silhouette dla obiektu i według poniższego wzoru:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Warto zauważyć, że powyższa wartość jest z zakresu $[-1, 1]$. Wartość silhouette dla poszczególnych klastrów jest średnią arytmetyczną silhouette dla każdego obiektu wewnątrz danego klastra. Globalna wartość silhouette wyznaczana jest jako średnia wartość silhouette dla każdego z klastrów. Im większa wartość globalna tym lepsza klasteryzacja.

1.3.3. Dunn

Miara dunn określa stopień kompaktowości klastrów. Oblicza się ją za pomocą dwóch wartości:

- d_{min} - minimalna odległość między elementami z dwóch różnych klastrów.
- d_{max} - maksymalna odległość między elementami z tego samego klastra.

Następnie należy zastosować poniższy wzór:

$$dunn(\Omega) = \frac{d_{min}}{d_{max}}$$

Warto zauważyć, że miara ta jest dodatnia. Im większa wartość tym lepiej dokonana klasteryzacja.

1.3.4. Davies-Bouldin

Miara davies-bouldin opiera się na pomiarach barycentrum klastrów (środków ciężkości). Określa ono podobieństwo pomiędzy wyznaczonymi grupami, im mniejsza wartość tym lepsza klasteryzacja. Miara jest dodatnia.

Rozdział 2

Dane testowe - charakterystyka

2.1. Seeds

2.1.1. Opis

Zbiór zawiera dane dotyczące charakterystki ziaren. Zawiera 3 klasy ziaren: Kama, Rosa and Canadian, po 70 rekordów każdy. Każde z ziaren opisane jest przez 7 atrybutów rzeczywistych. Łącznie 210 rekordów. Każde z ziaren jest opisane, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.1.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.1

Tab. 2.1: Dystrybucja klas w zbiorze seeds.

NAZWA	LICZNOŚĆ
Kama	70
Rosa	70
Canadian	70

2.1.3. Atrybuty

Poniżej 2.2 przedstawiono opis poszczególnych atrybutów.

Tab. 2.2: Opis atrybutów w zbiorze seeds.

NAZWA	RODZAJ	OPIS
Wielkość	REAL	area A
Perymetr	REAL	perimeter P
Kompaktowość	REAL	compactness C = $4 * pi * A/P^2$
Długość	REAL	length of kernel
Szerokość	REAL	width of kernel
Współczynnik asymetrii	REAL	asymmetry coefficient
Długość rowka	REAL	length of kernel groove

2.2. Ecoli

2.2.1. Opis

Zbiór zawiera dane dotyczące miejsca lokalizacji białek. Zawiera 8 klas lokalizacji przedstawione w tabeli 2.3. Każda z lokalizacji opisana jest przez 5 atrybutów rzeczywistych, 2 binarne, 1 kategoryczny, z czego ostatni jest nazwą konkretnej sekwencji (unikalna dla każdego rekordu). Łącznie 336 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji.

2.2.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.3

Tab. 2.3: Dystrybucja klas w zbiorze ecoli.

NAZWA	LICZNOŚĆ	OPIS
cp	143	cytoplasm
im	77	inner membrane without signal sequence
pp	52	perisplasm
imU	35	inner membrane, uncleavable signal sequence
om	20	outer membrane
omL	5	outer membrane lipoprotein
imL	2	inner membrane lipoprotein
imS	2	inner membrane, cleavable signal sequence

2.2.3. Atrybuty

Poniżej 2.4 przedstawiono opis poszczególnych atrybutów.

Tab. 2.4: Opis atrybutów w zbiorze seeds.

NAZWA	RODZAJ
name	UNIQUE
mcg	REAL
gvh	REAL
lip	BINARY
chg	BINARY
aac	REAL
alm2	REAL
alm1	REAL

2.3. User Knwoledge Modelling

2.3.1. Klasy

Zbiór zawiera dane dotyczące koorelacji między wynikami testów badanych osób, a czasem spędzonym na naukę. Zawiera 4 klasy oznaczające wynik egzaminu, przedstawione w tabeli 2.5. Każda z lokalizacji opisana jest przez 5 atrybutów rzeczywistych. Łącznie 403 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.3.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.5

Tab. 2.5: Dystrybucja klas w zbiorze seeds.

NAZWA	LICZNOŚĆ
very-low	50
low	129
middle	122
high	130

2.3.3. Atrybuty

Poniżej 2.8 przedstawiono opis poszczególnych atrybutów.

Tab. 2.6: Opis atrybutów w zbiorze ukm.

NAZWA	RODZAJ	OPIS
STG	REAL	Stopień poświęcenia czasu uczenia na główny cel
SCG	REAL	Stopień powtarzania informacji o głównym celu
STR	REAL	Stopień poświęcenia czasu uczenia na elementy powiązane z głównym celem
LPR	REAL	Wynik egzaminu powiązanego z głównym celem
PEG	REAL	Wynik egzaminu z głównym celem

2.4. Customers

2.4.1. Klasy

Zbiór zawiera dane dotyczące koorelacji między wydatkami na poszcczególne artykuły spożywcze w zależności od regionu w Portugalii. Zawiera 3 klasy oznaczające region z którego pochodzi dana statystyka (tabela 2.7). Każdy z regionów opisany jest przez 6 atrybutów rzeczywistych oznaczających wydatki na poszczególne artykuły. Łącznie 440 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.4.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.7

Tab. 2.7: Dystrybucja klas w zbiorze customers.

NAZWA	LICZNOŚĆ
Lisbon	77
Porto	47
Other	316

2.4.3. Atrybuty

Poniżej 2.8 przedstawiono opis poszczególnych atrybutów.

Tab. 2.8: Opis atrybutów w zbiorze ukm.

NAZWA	RODZAJ	OPIS
FRESH	REAL	Roczne wydatki na świeże produkty
MILK	REAL	Roczne wydatki na produkty mleczne
GROCERY	REAL	Roczne wydatki na jedzenie
FROZEN	REAL	Roczne wydatki na mrożonki
DETERGENTS_PAPER	REAL	Roczne wydatki na detergenty i artykuły papiernicze
DELICATESSEN	REAL	Roczne wydatki na delikatesy

Rozdział 3

Wyniki eksperymentów

3.1. Porównanie klasteryzacji względem ilości atrybutów

3.1.1. Zbiór seeds

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości atrybutów (kolumn).

Tab. 3.1: Wyniki klasteryzacji względem atrybutów. Zbiór seeds.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
seeds.csv	pam	euclidean	7	3	0,906	0,355	0,056
seeds.csv	kmeans	Forgy	7	3	0,778	0,397	0,069
seeds.csv	pam	euclidean	6	3	0,817	0,351	0,061
seeds.csv	kmeans	Forgy	6	3	0,694	0,396	0,062
seeds.csv	pam	euclidean	5	3	1,126	0,273	0,040
seeds.csv	kmeans	Forgy	5	3	0,878	0,335	0,042
seeds.csv	pam	euclidean	4	3	1,292	0,268	0,043
seeds.csv	kmeans	Forgy	4	3	0,906	0,332	0,045
seeds.csv	pam	euclidean	3	3	1,286	0,265	0,041
seeds.csv	kmeans	Forgy	3	3	0,963	0,343	0,043
seeds.csv	pam	euclidean	2	3	1,195	0,270	0,041
seeds.csv	kmeans	Forgy	2	3	1,085	0,333	0,041
seeds.csv	pam	euclidean	1	3	0,967	0,274	0,045
seeds.csv	kmeans	Forgy	1	3	0,955	0,334	0,047

3.1.2. Zbiór ecoli

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości atrybutów (kolumn).

Tab. 3.2: Wyniki klasteryzacji względem atrybutów. Zbiór ecoli.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
ecoli.csv	kmeans	Forgy	7	8	1,031	0,230	0,055
ecoli.csv	pam	euclidean	7	8	2,203	0,156	0,056
ecoli.csv	kmeans	Forgy	6	8	1,176	0,192	0,049
ecoli.csv	pam	euclidean	6	8	1,011	0,286	0,064
ecoli.csv	pam	euclidean	5	8	1,291	0,213	0,051
ecoli.csv	kmeans	Forgy	5	8	1,767	0,086	0,038
ecoli.csv	kmeans	Forgy	4	8	2,410	0,043	0,042
ecoli.csv	pam	euclidean	4	8	2,133	0,119	0,045
ecoli.csv	kmeans	Forgy	3	8	2,192	0,051	0,035
ecoli.csv	pam	euclidean	3	8	1,444	0,130	0,045
ecoli.csv	kmeans	Forgy	2	8	2,472	0,009	0,039
ecoli.csv	pam	euclidean	2	8	2,441	0,090	0,043
ecoli.csv	kmeans	Forgy	1	8	4,933	-0,025	0,035
ecoli.csv	pam	euclidean	1	8	2,506	0,090	0,042

3.1.3. Zbiór ukm

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości atrybutów(kolumn).

Tab. 3.3: Wyniki klasteryzacji względem atrybutów. Zbiór ukm.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
ukm.csv	pam	euclidean	5	4	1,45	0,17	0,08
ukm.csv	kmeans	Forgy	5	4	1,48	0,18	0,09
ukm.csv	pam	euclidean	1	4	4,06	0,02	0,06
ukm.csv	pam	euclidean	4	4	1,73	0,13	0,08
ukm.csv	kmeans	Forgy	4	4	1,64	0,15	0,08
ukm.csv	kmeans	Forgy	1	4	4,35	0,03	0,06
ukm.csv	pam	euclidean	3	4	1,91	0,10	0,06
ukm.csv	kmeans	Forgy	3	4	1,89	0,12	0,07
ukm.csv	pam	euclidean	2	4	2,84	0,04	0,06
ukm.csv	kmeans	Forgy	2	4	2,86	0,06	0,06

3.1.4. Zbiór customers

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości atrybutów(kolumn).

Tab. 3.4: Wyniki klasteryzacji względem atrybutów. Zbiór customers.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
customers.csv	kmeans	Forgy	6	3	0,906	0,272	0,017
customers.csv	pam	euclidean	6	3	0,966	0,209	0,012
customers.csv	pam	euclidean	5	3	0,893	0,202	0,013
customers.csv	kmeans	Forgy	5	3	0,925	0,267	0,015
customers.csv	pam	euclidean	4	3	0,987	0,194	0,012
customers.csv	kmeans	Forgy	4	3	0,851	0,273	0,015
customers.csv	kmeans	Forgy	3	3	0,949	0,275	0,014
customers.csv	pam	euclidean	3	3	0,982	0,203	0,011
customers.csv	pam	euclidean	2	3	1,221	0,138	0,010
customers.csv	kmeans	Forgy	2	3	1,012	0,223	0,012
customers.csv	kmeans	Forgy	1	3	1,700	0,177	0,011
customers.csv	pam	euclidean	1	3	2,052	0,092	0,008

3.2. Porównanie klasteryzacji względem ilości klastrów

3.2.1. Zbiór seeds

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości klastrów na jakie miał zostać podzielony zbiór danych.

Tab. 3.5: Wyniki klasteryzacji względem klastrów. Zbiór seeds.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
seeds.csv	kmeans	Forgy	7	10	1,831	0,210	0,051
seeds.csv	pam	euclidean	7	10	2,433	0,118	0,045
seeds.csv	kmeans	Forgy	7	9	1,112	0,296	0,058
seeds.csv	pam	euclidean	7	9	1,930	0,179	0,048
seeds.csv	kmeans	Forgy	7	8	0,797	0,329	0,068
seeds.csv	pam	euclidean	7	8	1,569	0,230	0,050
seeds.csv	kmeans	Forgy	7	7	0,820	0,325	0,067
seeds.csv	pam	euclidean	7	7	1,219	0,282	0,055
seeds.csv	kmeans	Forgy	7	6	0,809	0,324	0,069
seeds.csv	pam	euclidean	7	6	0,998	0,312	0,060
seeds.csv	pam	euclidean	7	5	0,947	0,318	0,053
seeds.csv	kmeans	Forgy	7	5	0,814	0,330	0,069
seeds.csv	pam	euclidean	7	4	0,935	0,325	0,052
seeds.csv	kmeans	Forgy	7	4	0,814	0,354	0,069
seeds.csv	pam	euclidean	7	3	0,906	0,355	0,056
seeds.csv	kmeans	Forgy	7	3	0,778	0,397	0,069
seeds.csv	pam	euclidean	7	2	0,819	0,386	0,054
seeds.csv	kmeans	Forgy	7	2	0,726	0,448	0,058

3.2.2. Zbiór ecoli

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości klastrów na jakie miał zostać podzielony zbiór danych.

Tab. 3.6: Wyniki klasteryzacji względem klastrów. Zbiór ecoli.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
ecoli.csv	kmeans	Forgy	7	10	2,687	0,119	0,057
ecoli.csv	pam	euclidean	7	10	3,390	0,078	0,058
ecoli.csv	kmeans	Forgy	7	9	1,616	0,176	0,057
ecoli.csv	pam	euclidean	7	9	2,805	0,115	0,058
ecoli.csv	kmeans	Forgy	7	8	1,031	0,230	0,055
ecoli.csv	pam	euclidean	7	8	2,203	0,156	0,056
ecoli.csv	kmeans	Forgy	7	7	1,150	0,242	0,056
ecoli.csv	pam	euclidean	7	7	1,428	0,198	0,057
ecoli.csv	kmeans	Forgy	7	6	1,341	0,238	0,058
ecoli.csv	pam	euclidean	7	6	1,102	0,232	0,058
ecoli.csv	pam	euclidean	7	5	1,017	0,249	0,056
ecoli.csv	kmeans	Forgy	7	5	1,397	0,229	0,056
ecoli.csv	pam	euclidean	7	4	1,024	0,262	0,057
ecoli.csv	kmeans	Forgy	7	4	1,258	0,265	0,069
ecoli.csv	pam	euclidean	7	3	1,019	0,277	0,061
ecoli.csv	kmeans	Forgy	7	3	1,179	0,306	0,078
ecoli.csv	pam	euclidean	7	2	1,003	0,308	0,070
ecoli.csv	kmeans	Forgy	7	2	1,000	0,375	0,092

3.2.3. Zbiór ukm

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości klastrów na jakie miał zostać podzielony zbiór danych.

Tab. 3.7: Wyniki klasteryzacji względem klastrów. Zbiór ukm.

name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
ukm.csv	pam	euclidean	5	10	1,5392363	0,1636159	0,0184714
ukm.csv	kmeans	Forgy	5	10	1,2144818	0,2719116	0,0294932
ukm.csv	pam	euclidean	5	9	1,498921	0,1794329	0,0285728
ukm.csv	kmeans	Forgy	5	9	1,1742246	0,2545442	0,0588987
ukm.csv	pam	euclidean	5	8	1,4642177	0,1902306	0,0435276
ukm.csv	kmeans	Forgy	5	8	1,2787643	0,1889988	0,086558
ukm.csv	pam	euclidean	5	7	1,3899226	0,1941117	0,0557502
ukm.csv	kmeans	Forgy	5	7	1,3544563	0,1858345	0,0871468
ukm.csv	pam	euclidean	5	6	1,3775152	0,1798164	0,0700398
ukm.csv	kmeans	Forgy	5	6	1,3759643	0,1837686	0,0900128
ukm.csv	pam	euclidean	5	5	1,4218182	0,1659925	0,075158
ukm.csv	kmeans	Forgy	5	5	1,4128306	0,1827811	0,0922943
ukm.csv	pam	euclidean	5	4	1,4488258	0,1660448	0,0787899
ukm.csv	kmeans	Forgy	5	4	1,4793784	0,1811719	0,0936553
ukm.csv	pam	euclidean	5	3	1,5040453	0,168982	0,075142
ukm.csv	kmeans	Forgy	5	3	1,5245143	0,1816338	0,0953541
ukm.csv	pam	euclidean	5	2	1,5716208	0,1729939	0,0744757
ukm.csv	kmeans	Forgy	5	2	1,6446009	0,1856208	0,0880215

3.2.4. Zbiór customers

Poniżej przedstawiono wyniki klasteryzacji względem różnej ilości klastrów na jakie miał zostać podzielony zbiór danych.

Tab. 3.8: Wyniki klasteryzacji względem klastrów. Zbiór customers.

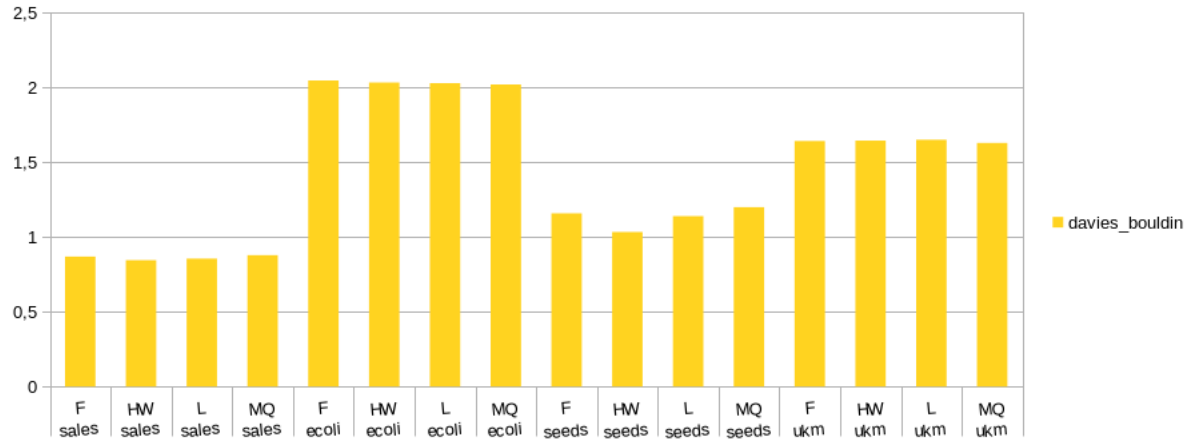
name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
customers.csv	kmeans	Forgy	6	10	0,848	0,361	0,048
customers.csv	pam	euclidean	6	10	0,087	0,018	0,001
customers.csv	kmeans	Forgy	6	9	0,830	0,314	0,034
customers.csv	pam	euclidean	6	9	0,286	0,056	0,003
customers.csv	kmeans	Forgy	6	8	0,971	0,200	0,013
customers.csv	pam	euclidean	6	8	0,426	0,097	0,005
customers.csv	kmeans	Forgy	6	7	1,012	0,214	0,014
customers.csv	pam	euclidean	6	7	0,636	0,136	0,006
customers.csv	kmeans	Forgy	6	6	0,984	0,215	0,014
customers.csv	pam	euclidean	6	6	0,835	0,175	0,009
customers.csv	kmeans	Forgy	6	5	0,937	0,218	0,015
customers.csv	pam	euclidean	6	5	0,905	0,199	0,010
customers.csv	kmeans	Forgy	6	4	0,891	0,241	0,015
customers.csv	pam	euclidean	6	4	0,896	0,206	0,011
customers.csv	kmeans	Forgy	6	3	0,906	0,272	0,017
customers.csv	pam	euclidean	6	3	0,966	0,209	0,012
customers.csv	kmeans	Forgy	6	2	0,968	0,292	0,017
customers.csv	pam	euclidean	6	2	0,985	0,225	0,013

3.3. Porównanie metod kmeans

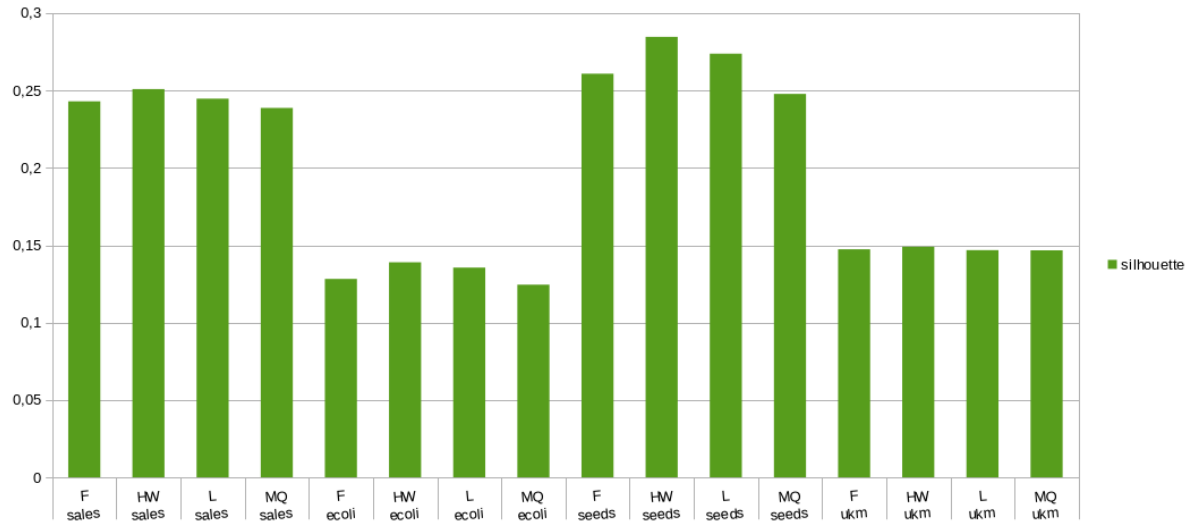
Poniżej przedstawiono wyniki klasteryzacji algorytmem kmeans przy użyciu różnych metod - MacQueen, Lloyd, Hartigan-Wong, Forgy.

Tab. 3.9: Porównanie różnych metod w kmeans.

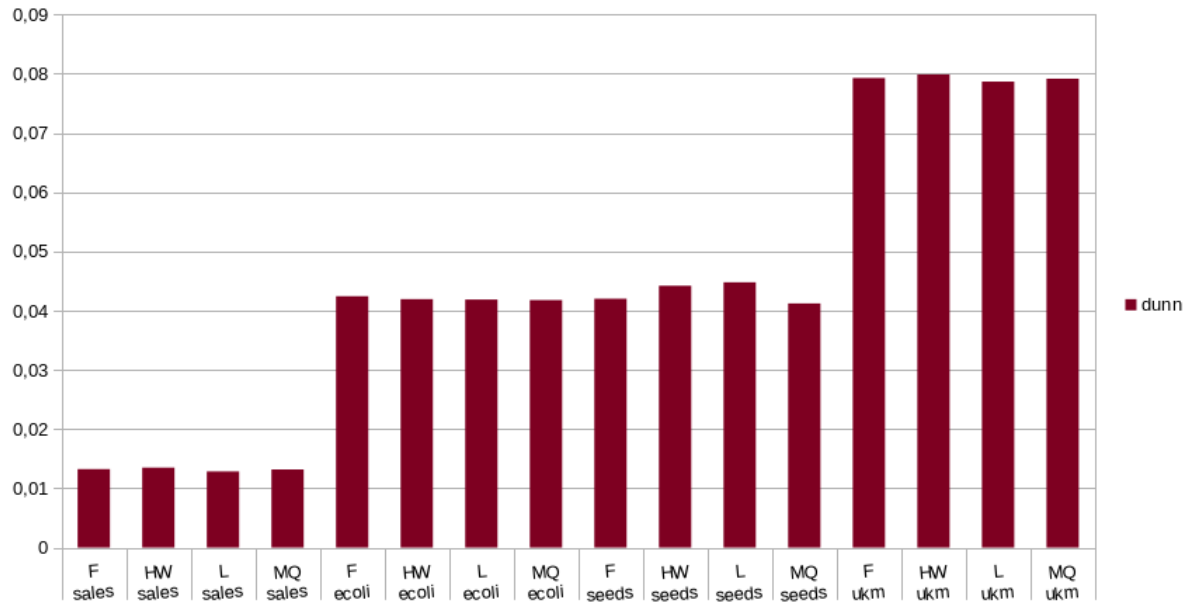
name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
sales	kmeans	Forgy	4	4	0,869	0,243	0,013
sales	kmeans	Hartigan-Wong	4	4	0,845	0,251	0,013
sales	kmeans	Lloyd	4	4	0,856	0,245	0,013
sales	kmeans	MacQueen	4	4	0,878	0,239	0,013
ecoli	kmeans	Forgy	4	4	2,047	0,128	0,042
ecoli	kmeans	Hartigan-Wong	4	4	2,033	0,139	0,042
ecoli	kmeans	Lloyd	4	4	2,029	0,136	0,042
ecoli	kmeans	MacQueen	4	4	2,020	0,125	0,042
seeds	kmeans	Forgy	4	4	1,159	0,261	0,042
seeds	kmeans	Hartigan-Wong	4	4	1,034	0,285	0,044
seeds	kmeans	Lloyd	4	4	1,140	0,274	0,045
seeds	kmeans	MacQueen	4	4	1,200	0,248	0,041
ukm	kmeans	Forgy	4	4	1,642	0,147	0,079
ukm	kmeans	Hartigan-Wong	4	4	1,645	0,149	0,080
ukm	kmeans	Lloyd	4	4	1,651	0,147	0,079
ukm	kmeans	MacQueen	4	4	1,629	0,147	0,079



Rys. 3.1: Porównanie różnych metod w kmeans. Miara Davies-Bouldin Index.



Rys. 3.2: Porównanie różnych metod w kmeans. Miara Silhouette.



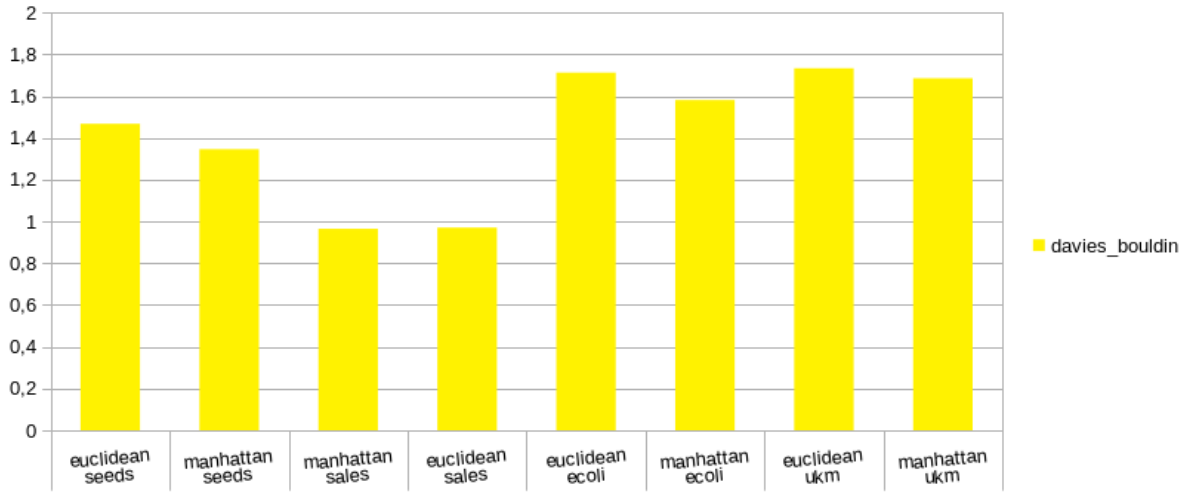
Rys. 3.3: Porównanie różnych metod w kmeans. Miara Dunn.

3.4. Porównanie metod pam

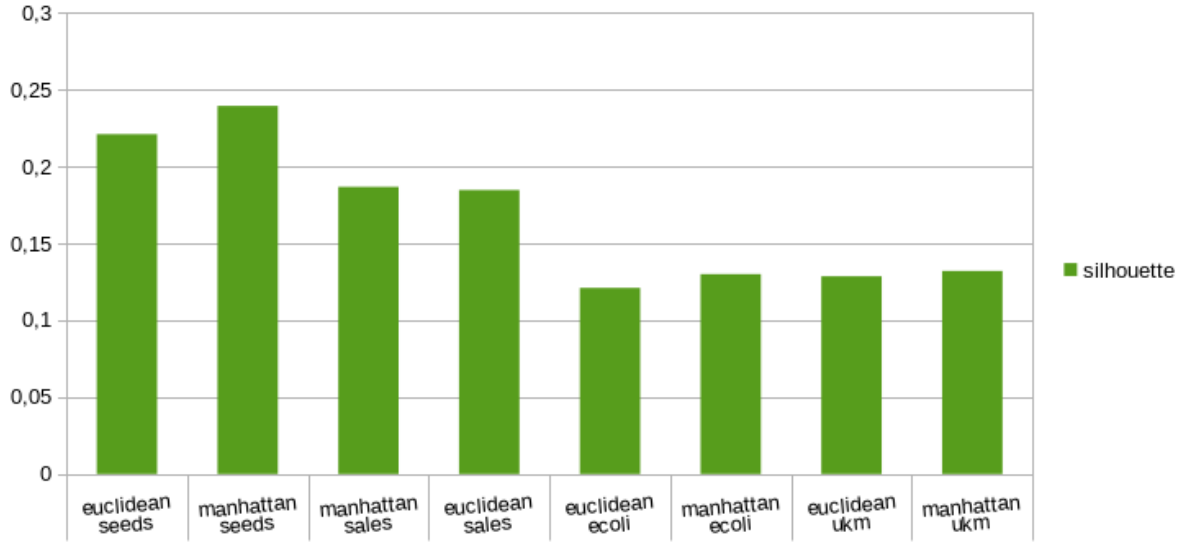
Poniżej przedstawiono wyniki klasteryzacji algorytmem kmeans przy użyciu różnych metryk - euklidesowa i manhattan.

Tab. 3.10: Porównanie różnych metryk w pam.

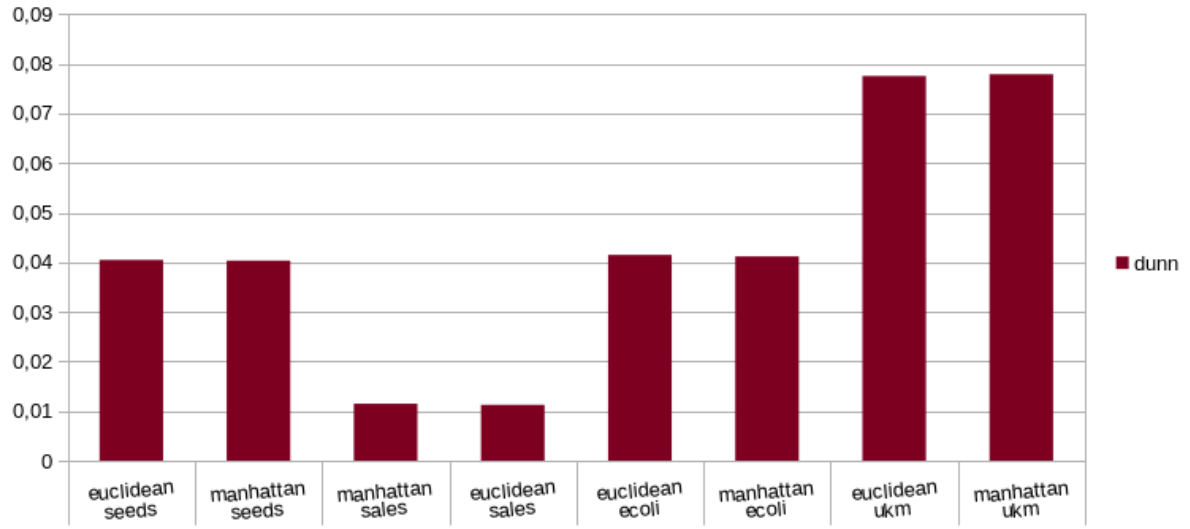
name	algo	param	columns	clusters	davies_bouldin	silhouette	dunn
seeds	pam	euclidean	4	4	1,468	0,221	0,040
seeds	pam	manhattan	4	4	1,346	0,239	0,040
sales	pam	manhattan	4	4	0,965	0,187	0,011
sales	pam	euclidean	4	4	0,971	0,185	0,011
ecoli	pam	euclidean	4	4	1,713	0,121	0,041
ecoli	pam	manhattan	4	4	1,582	0,130	0,041
ukm	pam	euclidean	4	4	1,734	0,129	0,078
ukm	pam	manhattan	4	4	1,686	0,132	0,078



Rys. 3.4: Porównanie różnych metryk w pam. Miara Davies-Bouldin Index.



Rys. 3.5: Porównanie różnych metryk w pam. Miara Silhouette.



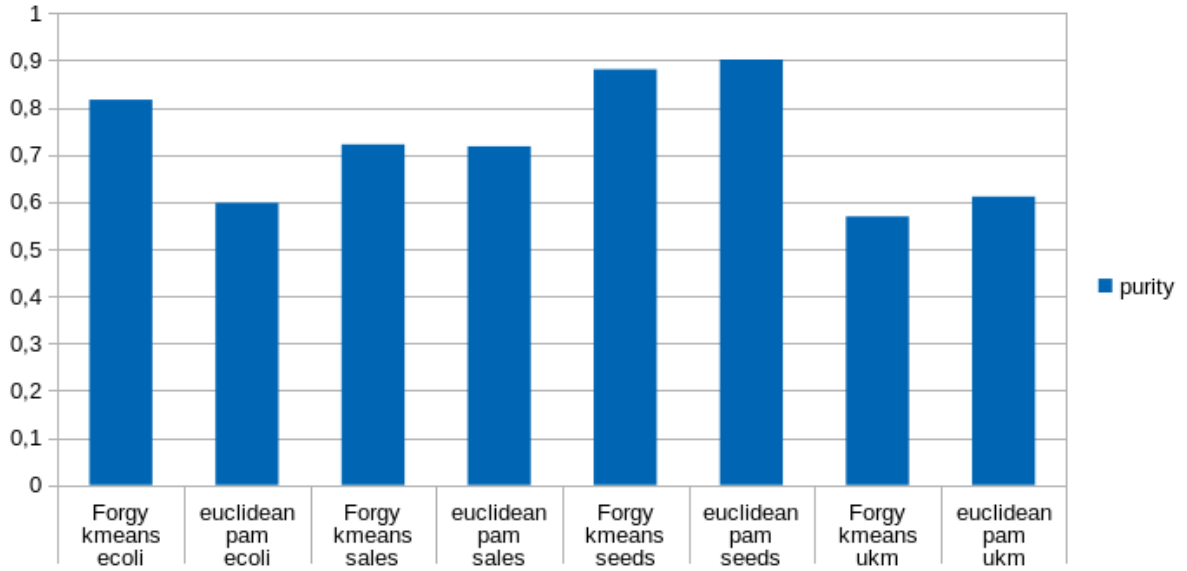
Rys. 3.6: Porównanie różnych metryk w pam. Miara Dunn.

3.5. Porównanie dopasowania klasteryzacji - miara purity

Poniżej przedstawiono wyniki dopasowania klasteryzacji do obecnych klas w poszczególnych zbiorach za pomocą miary purity.

Tab. 3.11: Porównanie dopasowania.

name	algo	param	columns	clusters	purity
ecoli	kmeans	Forgy	7	8	0,817
ecoli	pam	euclidean	7	8	0,598
sales	kmeans	Forgy	6	3	0,722
sales	pam	euclidean	6	3	0,718
seeds	kmeans	Forgy	7	3	0,881
seeds	pam	euclidean	7	3	0,902
ukm	kmeans	Forgy	5	4	0,569
ukm	pam	euclidean	5	4	0,611



Rys. 3.7: Porównanie dopasowania - purity.

Rozdział 4

Wnioski

4.1. Ilość atrybutów w klasteryzacji

Dla zbioru seeds najlepsze wyniki klasteryzacji osiągnięto dla wszystkich, bądź prawie wszystkich dostępnych atrybutów. W przypadku zbiorów ecoli oraz ukm osiągnięto znacznie słabszą klasteryzację względem zbioru seeds (bardzo niski wskaźnik silhouette). Jednakże i tu sprawdziła się zależność dotycząca ilości atrybutów - im więcej tym lepsza klasteryzacja. Dla zbioru porównawczego customers, osiągnięto podobne wyniki. Warto zauważyć, że indeks Dunn dla tego zbioru osiągnął bardzo niską wartość, co świadczyć może o małym zróżnicowaniu poszczególnych danych w zbiorze. Niezależnie od zbioru obie metody - PAM i K-means osiągnęły tutaj porównywalną wartość.

4.2. Ilość klastrów

Jakość klasteryzacji w zależności od metody i ilości klastrów charakteryzowała się podobną zmiennością. Dla małej ilości klastrów obie metody uzyskiwały podobne wyniki. Wraz ze wzrostem ilości klastrów metoda k-means uzyskała lepsze wyniki we wszystkich miarach niż metoda PAM niezależnie od zbioru. Warto zauważyć, że w przypadku metody PAM, najlepsze wyniki odnotowano dla małych klastrów, a dla k-means odwrotnie - najlepsze wyniki dla największej ilości klastrów.

4.3. Parametry kmeans

W przypadku różnych odmian algorytmu k-means, odnotowano bardzo zbliżone do siebie wyniki dla poszczególnych zbiorów. Algorytm Hartigan-Wong odnotował lekką przewagę (około 10%) nad innymi algorytmami w przypadku zbioru seeds.

4.4. Parametry pam

Zastosowanie różnych metryk w algorytmie PAM spowodowało zróżnicowanie indeksów silhouette i davies-bouldin dla poszczególnych zbiorów. Metryka Manhattan okazała się nieznacznie lepsza w zbiorach seeds, ecoli oraz ukm. Dla zbioru customers nie odnotowano znaczących różnic.

4.5. Dopasowanie - purity

Najlepsze dopasowanie klastrów do aktualnych klas odnotowano dla zbioru seeds i wyniosło ono blisko 0.9. Najsłabsze zaś dla zbioru ukm - wartość oscylowała w granicach 0.6. Dla wszystkich zbiorów, oprócz ecoli, obie metody (k-means i pam) charakteryzowały się podobnym dopasowaniem. W przypadku zbioru ecoli różnica pomiędzy metodami wyniosła aż 0.2 na korzyść k-means.

Warto również zauważyć, żę miara purity jest miarą, która może wprowadzać w błąd, gdyż dla różnych rozmieszczeń klastrów miara ta może osiągnąć tę samą wartość. Przykładowo rozpatrzmy następujące dwa wyniki klasteryzacji:

Tab. 4.1: Przykładowe rozmieszczenie klastrów.

	A	B	C
C_1	90	8	5
C_2	3	90	5
C_3	7	2	90

	A	B	C
C_1	90	90	90
C_2	3	8	5
C_3	7	2	5

W obu przypadkach purity wynosi 0.9. Jednakże dopasowanie w pierwszym przypadku jest znacznie lepsze gdyż, poszczególne klastry posiadają zbliżoną ilość obiektów i w podobny sposób mapują się na klasy (po 100 elementów w klasie). Dla drugiego przypadku klaster C_1 jest znacznie większy od pozostałych i posiada cechy ze wszystkich klas, więc w tym przypadku klastry bardzo słabo odwzorowują poszczególne klasy.