Human Activity Recognition with Smartphones - Klastrowanie

Adrian Kamiński i Michał Komorowski

28 maja 2021

1 Wstęp

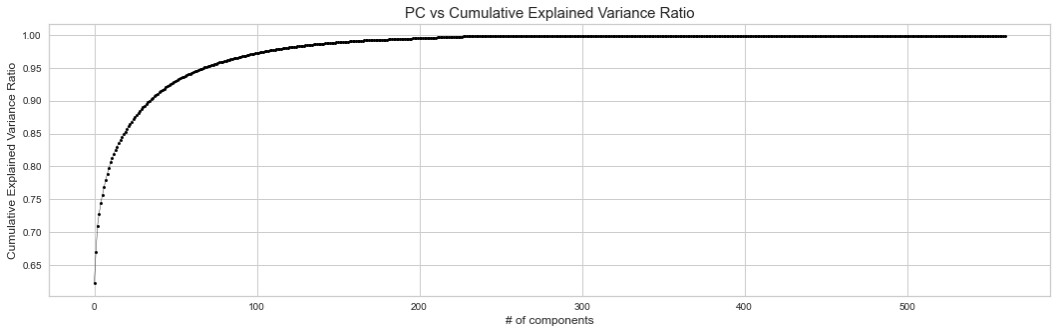
W drugim projekcie na przedmiocie *Wstęp do Uczenia maszynowego* zmierzyliśmy się z problemem klastrowania. W odróżnieniu od klasyfikacji pozbywamy się tu zmiennej celu i próbujemy znaleźć w zbiorze zależności i przypisać obserwacje, które są do siebie odpowiednio podobne do klastrów.

2 Opis zbioru danych

Zbiór danych, którym się zajmowaliśmy zawiera odczyty z różnych czytników ze smartfona takich jak akcelerometr czy żyroskop. Każdy rodzaj pomiaru ma wiele statystyk takich jak średnia, mediana czy odchylenie standardowe. Dodatkowo, niektóre pomiary są rozdzielone na trzy wymiary X, Y oraz Z. Przez to, zbiór jest dosyć duży, bo zawiera aż 561 kolumn. Sporo z nich jest współliniowych.

3 PCA

Aby zredukować rozmiar danych skorzystaliśmy z PCA. Spróbowaliśmy różnych liczb komponentów, mianowicie tak by zostało zachowane 95%, 90% oraz 80% wariancji.

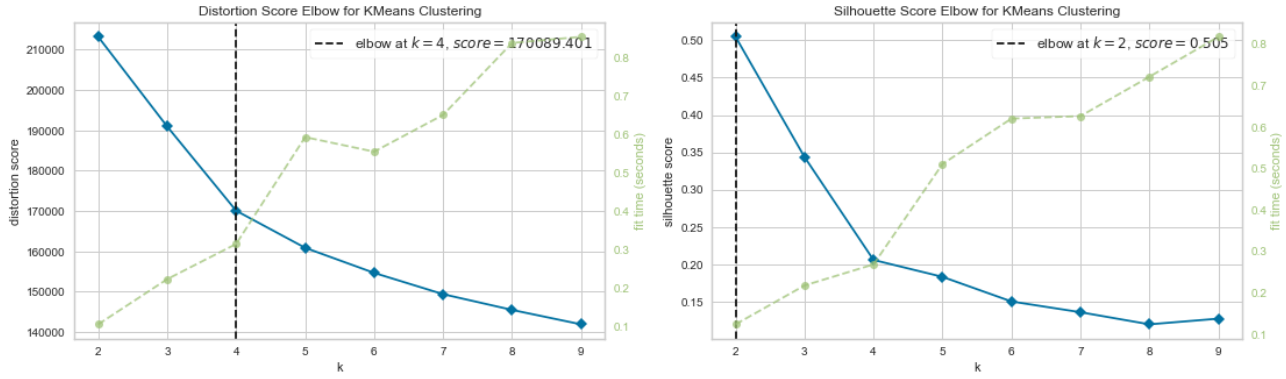


Dla konkretnych tych wartości otrzymaliśmy odpowiednio 69, 36 oraz 11 komponentów co znacząco zmniejsza rozmiar danych o ponad 500 kolumn.

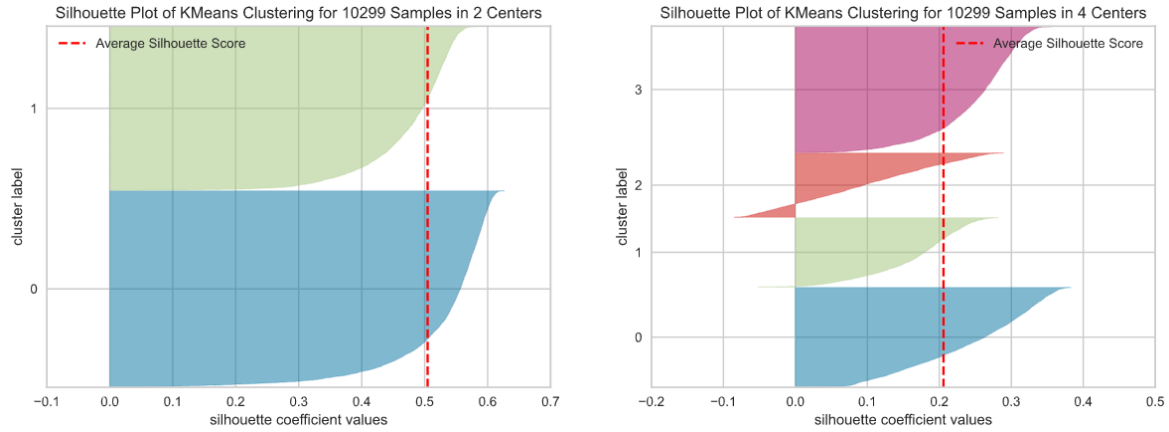
4 Modelowanie

4.1 Kmeans

Pierwszą metodą było K-średnich, posłużyliśmy się pakietem *yellowbrick* aby łatwo zwizualizować metodę łokcia i silhouette samples. Otrzymaliśmy tutaj dwie opcje dla optymalnej liczby klastrów, czyli 2 i 4. Powtórzyliśmy to dla trzech opcji PCA, jednak różnice były niewielkie. Można więc tu śmiało przyjąć dane z najmniejszą liczbą komponentów.



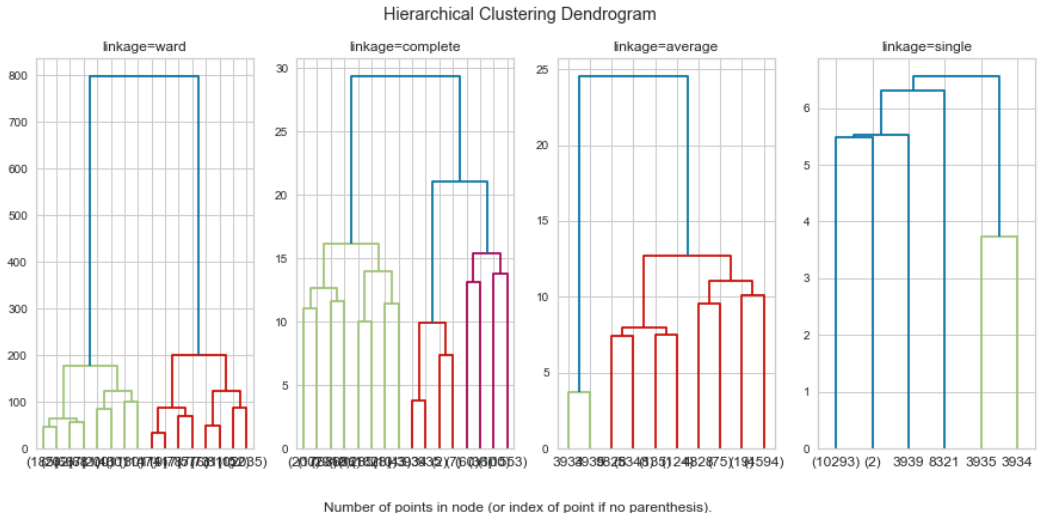
Rysunek 1: PCA n_components = 69 (95% EVR)



Rysunek 2: Silhouette plot n_components = 69 (95% EVR)

4.2 AgglomerativeClustering

Następnie skorzystaliśmy z klastrowania hierarchicznego. Skorzystaliśmy z 4 możliwych połączeń (linkage) i za pomocą dendrogramu wybraliśmy liczbę klastrów, czyli 2, 3, 3 i 2 odpowiednio dla linkage: ward, complete, average i single.



Rysunek 3: PCA n_components = 69 (95% EVR)

4.3 Inne modele

Próbowaliśmy innych też innych modeli. Pierwszym z nich był DBScan, jednak nic sensownego nie udało nam się osiągnąć, więc zrezygnowaliśmy z niego. Drugim było Gaussian Mixture Models i tutaj skorzystaliśmy podobnie jak w KMeans z 2, 3 oraz 4 klastrów.

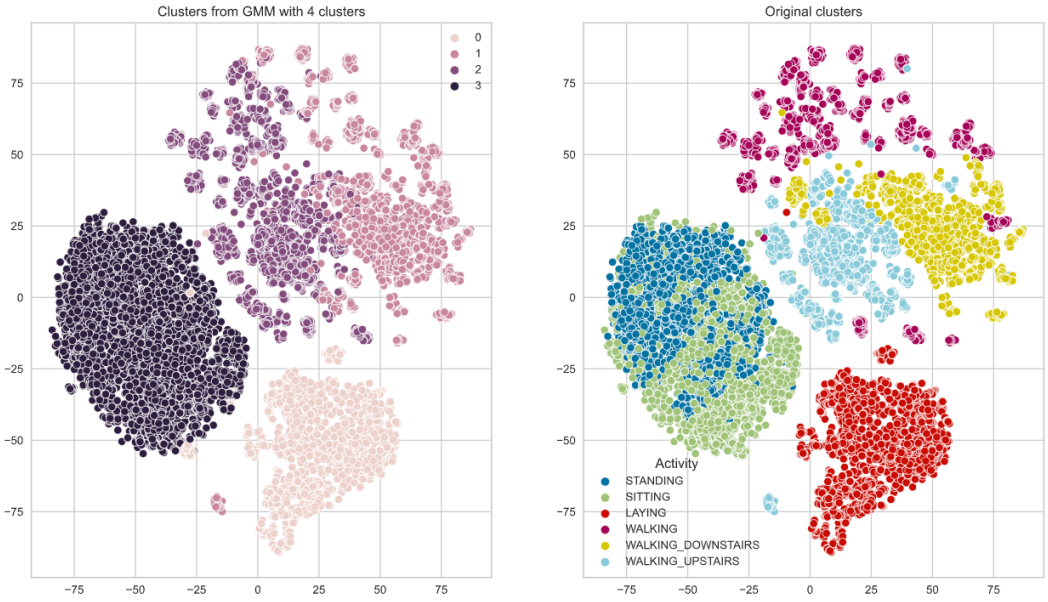
5 Podsumowanie

5.1 Wyniki

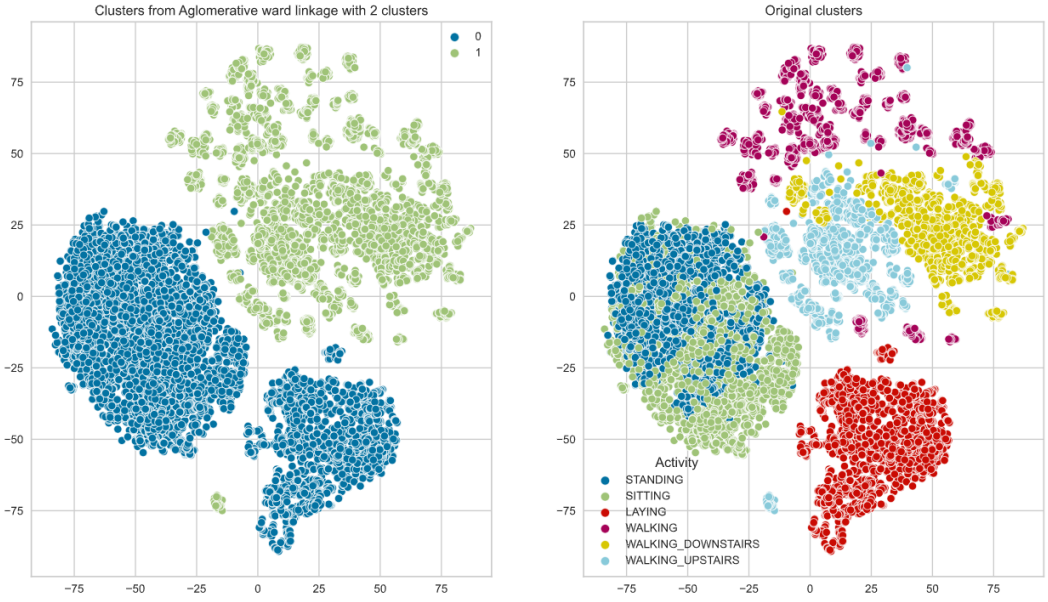
		n_clusters	silhouette	davies-bouldin	calinski-harabasz	accuracy
PCA 69	Kmeans	**2**	0.504670	0.792377	15345.115308	0.998544
components		**3**	0.343609	1.551199	9164.246446	0.877367
95%EVA		**4**	0.205838	1.892348	7279.926445	0.873871
	Gaussian Mixture Models	**2**	0.504237	0.792963	15306.177501	0.999417
		3	0.330909	1.504405	8944.451433	0.862802
		4	0.188806	1.778222	7104.790754	0.882707
	Agglomerative ward linkage	**2**	0.503997	0.793327	15282.235723	1.000000
	Agglomerative single linkage	**2**	0.594533	0.373060	21.974284	0.546558
	Agglomerative complete linkage	**3**	0.262647	0.978491	1998.961270	0.642781
	Agglomerative average linkage	**3**	0.500601	0.640262	7696.739841	0.861831
PCA 35	Kmeans	**2**	0.532132	0.736858	17576.884533	0.998544
components		**3**	0.368817	1.423686	10675.125972	0.877075
90%EVA		**4**	0.232227	1.727457	8654.706709	0.874454
	Gaussian Mixture Models	**2**	0.531416	0.737820	17501.922496	0.999515
		3	0.360221	1.370935	10506.923133	0.862899
		4	0.212970	1.600950	7321.631279	0.887950
	Agglomerative ward linkage	**2**	0.531389	0.737893	17498.609179	0.999223
	Agglomerative single linkage	**2**	0.593689	0.360654	21.945381	0.546558
	Agglomerative complete linkage	**3**	0.436064	1.128632	9732.182137	0.882222
	Agglomerative average linkage	**3**	0.527130	0.595946	8830.140990	0.862123
PCA 35	Kmeans	**2**	0.601402	0.609455	24688.620887	0.998544
components		**3**	0.438412	1.115910	15838.857531	0.877561
80%EVA		**4**	0.314958	1.332598	13804.966476	0.874357
	Gaussian Mixture Models	**2**	0.600527	0.610446	24555.060823	0.999320
		3	0.380586	1.414101	14223.265309	0.863288
		4	0.224425	1.702999	11267.956029	0.726672
	Agglomerative ward linkage	**2**	0.600490	0.610441	24552.392090	0.999417
	Agglomerative single linkage	**2**	0.612259	0.307086	23.581742	0.546558
	Agglomerative complete linkage	**3**	0.482192	0.631382	7596.823566	0.785707
	Agglomerative average linkage	**3**	0.569451	0.604285	12847.618517	0.866298

Tablica 1: Wyniki wszystkich modeli

5.2 Wizualizacja wybranych modeli



Rysunek 4: GMM 4 klastry



Rysunek 5: Aglomeracyjne 2 klastry

5.3 Interpretacja klastrów

W przypadku podziału na cztery klastry możemy wyróżnić aktywności: LAYING, STANDING i SITTING, WALKING WALKING_UPSTAIRS WALKING_DOWNSTAIRS.

Podział na dwa klastry to podział na dwa typy aktywności: dynamiczne (WALKING) oraz stacjonarne (STANDING, SITTING, LAYING)