Analiza skupień: wizualizacja

Mateusz Staniak, 7 V 2019

Problem

Dla danych wysokowymiarowych (a tak naprawdę już dla wymiaru większego od 2) wizualizacja pogrupowania danych jest trudna.

PCA

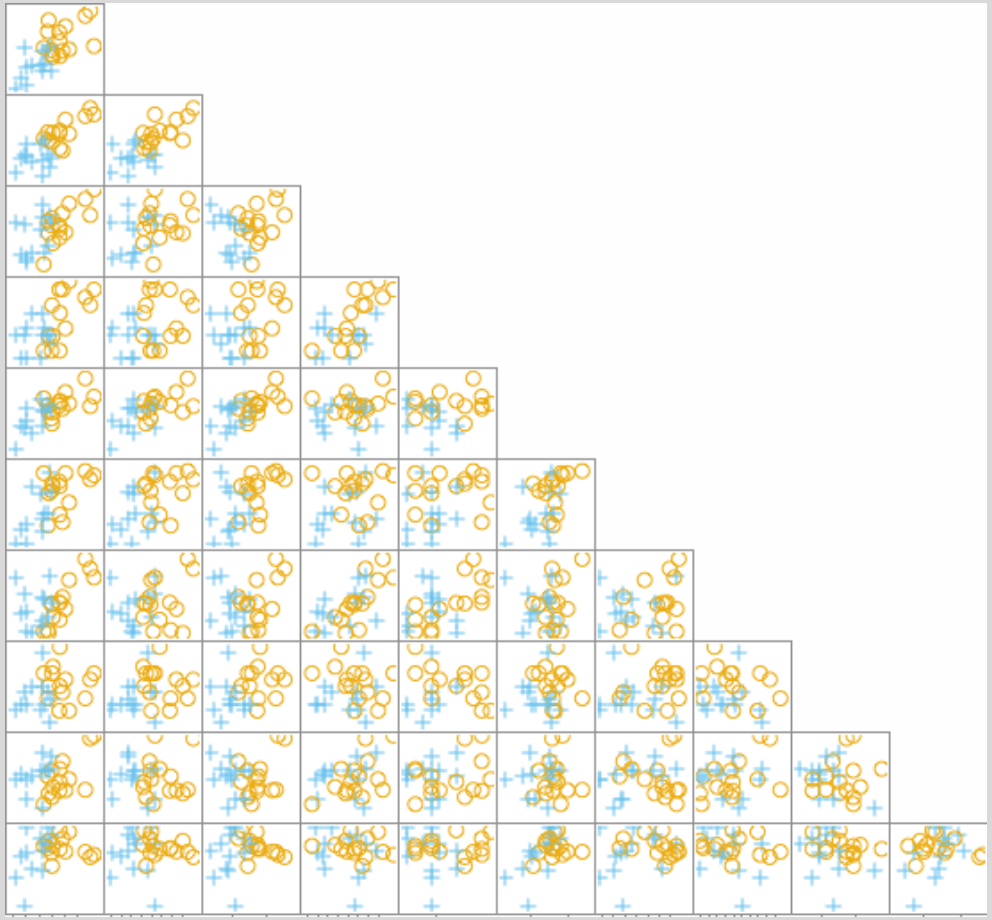
1. Obliczyć klastry dowolnym algorytmem.
2. Zredukować wymiar danych PCA.
3. Narysować dwie składowe główne z klastrami zaznaczonymi kolorem.

-> Uwaga: PCA daje pewne pojęcie o tym, jak wyglądają klastry. Ale składowe główne nie są dobrymi zmiennymi rozdzielającymi (PCA nie jest algorytmem klasteryzacji, pomaga tylko w wizualizacji)

https://rpkgs.datanovia.com/factoextra/reference/fviz_cluster.html

Alternatywa

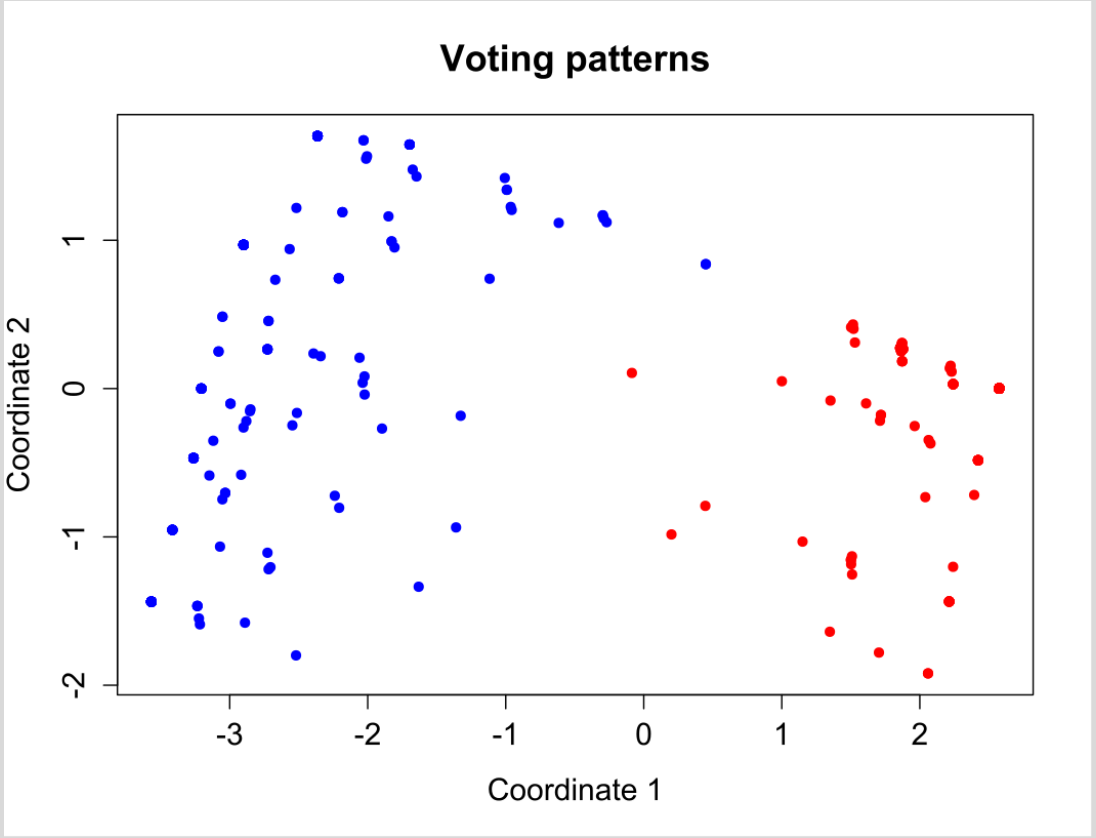
- Zamiast wizualizować składowe główne, możemy wizualizować klastry pomiędzy wszystkimi parami zmiennych (macierz wykresów rozrzutu + kolor oznaczający klaster)



https://stats.stackexchange.com/questions/52625/visually-plotting-multi-dimensional-cluster-data

Skalowanie wielowymiarowe (MDS)

- Metoda wizualizacji (nie-)podobieństwa między obserwacjami.
- W klasycznej wersji opiera się na PCA:
  1. Obliczyć macierz kwadratów odległości między obserwacjami.
  2. Zastosować PCA do tej macierzy.
- Istnieją uogólnienia tej metody.



https://en.wikipedia.org/wiki/Multidimensional_Scaling

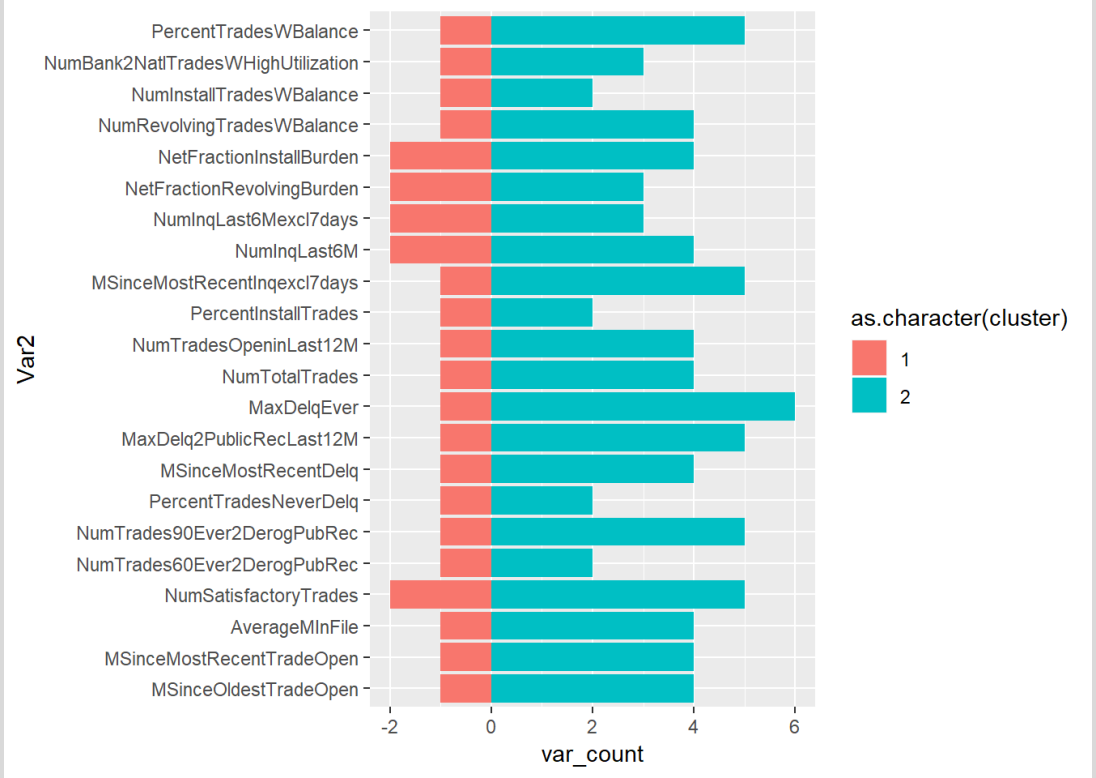
t-SNE

- Nieliniowa metoda redukcji wymiaru.
- Zachowuje bliskie punkty lepiej niż metody liniowe.
- Opiera się na minimalizacji różnic między prawdopodobieństwami „sąsiedztwa” między punktami w przestrzeni wysokowymiarowej i niskowymiarowej (tę, na którą „rzutujemy”).
- Szczegóły: https://lvdmaaten.github.io/tsne/
- Przydatne: https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/

https://lvdmaaten.github.io/tsne/examples/caltech101_tsne.jpg

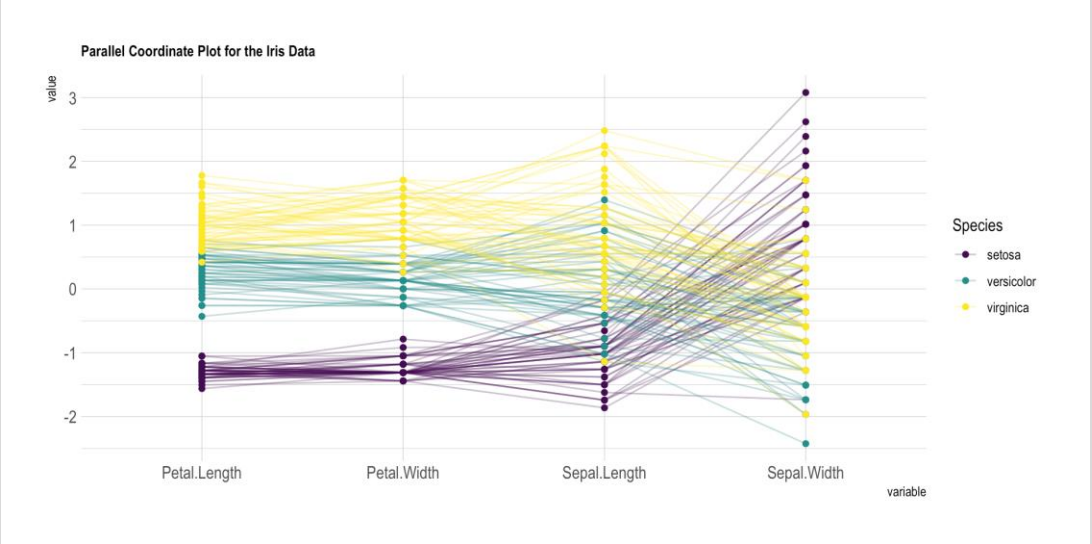
Porównania rozkładów cech pomiędzy klastrami

- Porównanie rozkładów cech pomaga zrozumieć różnice pomiędzy klastrami.
- Np. wykresy słupkowe typu „piraida wieku” (po prawej), histogramy, gęstości, wykresy słupkowe, wykresy skrzypcowe itd. Narysowane wg klastrów.



Parallel Coordinate Plot

- Parallel Coordinate Plot to wykres uogólniający dwuwymiarowy wykres rozrzutu. Osie są dodawane równolegle zamiast prostopadle (stąd nazwa…).
- Jednym z jego zastosowań jest wizualna identyfikacja klastrów.
- https://www.data-to-viz.com/graph/parallel.html
- https://amstat.tandfonline.com/doi/abs/10.1080/10618600.2012.694762?journalCode=ucgs20



https://www.data-to-viz.com/graph/parallel.html

Różne metody

- Block Clustering
- Data Image
- Generalized Association Plots (GAP)
- Klik: http://gap.stat.sinica.edu.tw/Talks/Hank-ClusterVisualization.pdf

Materiały

- https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/
- https://docs.google.com/viewer?url=http://www.schonlau.net/publication/04compstat_clustergram.pdf - clustergram do wizualizacji wyników klastrowania
- https://apps.dtic.mil/dtic/tr/fulltext/u2/a313545.pdf - wizualizacja wysokowymiarowych danych (Parallel Coordinate Plot + Tours)
- https://collabspace.cornell.edu/projects/visualizing-hi-dimensional-data-using-parallel-coordinates
- https://dicook.public.iastate.edu/JSS/paper/paper.html - „Calibrate Your Eyes to Recognize High-Dimensional Shapes from Low-Dimensional Projections”
- http://gap.stat.sinica.edu.tw/Talks/Hank-ClusterVisualization.pdf

Pakiety w R

- mixClustType (nowe pakiet do klasteryzacji metodą k prototypów dla danych mieszanych)
- cluster (typowy pakiet)
- GGAlly, Ggfortify (wielowymiarowe wykresy)
- https://rpkgs.datanovia.com/factoextra/index.html (wizualizacja + redukcja wymiaru)
- Różności: tsne, flexclust, flexmix, fpc, gplots, kohonen, mvtnorm, vcd.

(Narzędzia w Pythonie: TBA)

Sonda: co na zadanie domowe?

- Metody wizualizacji z prezentacji http://gap.stat.sinica.edu.tw/Talks/Hank-ClusterVisualization.pdf?
- Metody diagnostyki/walidacji (clustergram, metody z https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/#determining-the-optimal-number-of-clusters, wybór liczby klastrów)?
- t-SNE?
- Inne?

Zadanie na laboratorium

- Wybrać jeden z projektów.
- Zastosować metody wizualizacji ze slajdów do danych z wybranego projektu. (Znalezienie odpowiedniej biblioteki + wykonanie wykresów + wnioski -> przyda się do I części projektu).