UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Simple & Scalable Sparse K-means clustering with Feature Ranking

IE 529: Course Project

Team members

Praveen Kumar M (Net ID: pkm4)

Lloyd Fernandes (Net ID: lloydf2)

Omkar Rajendra Mehta (Net ID: omehta2)

Vincent Hoff (Net ID: vhoff2)

December 08, 2021

CLUSTERING

- **Clustering** is one of the most widely used statistical techniques to group data into homogenous clusters, with each cluster sharing similar characteristics.

- ***K*-means** is a popular method for performing clustering owing to its simplicity, speed and familiarity. The objective is to reduce the within-cluster sums of squared losses (aka WCSS). The steps include:
  - Assigning each data point to its nearest cluster center
  - Updating cluster centers by mean of the data points of each cluster

PROBLEM AND MOTIVATION

- For high dimensional data :
  - Signal to noise ratio reduces with each additional feature. Clustering ability deteriorates
  - Computational and memory requirements increases

- Many algorithms have been developed to address these issues as k-means clustering does not work.

PAST WORK

- **Filter based methods**
  - Filters the right features which can improve clustering quality before applying clustering.
- **Wrapper based methods**
  - Select subsets of attributes and identify the best subset post clustering - based on certain cluster quality metrics.
  - **Clustering on Subset of attributes (COSA)** : It is a Wrapper based method where weights are assigned to important features which are higher compared to other features.
  - **Sparse k-means (SKM)** adds the l-1 norm of weights for calculating weights of the features using gradient descent and selecting features.

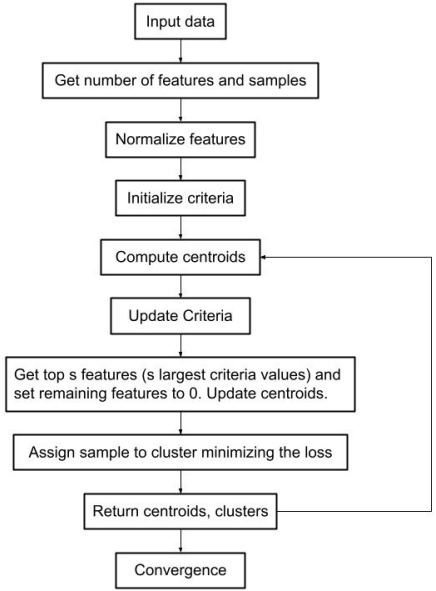
$$\sum_{m=1}^{p} w_m \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ijm} - \sum_{l=1}^{k} \frac{1}{|C_l|} \sum_{i,j \in C_l} d_{ijm} \right)$$

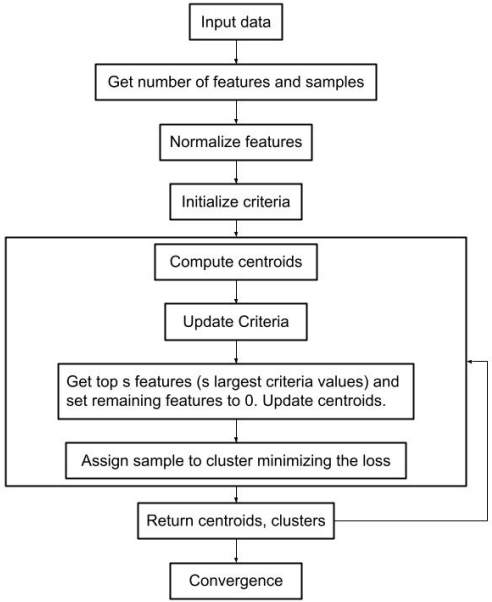
Sparse and Scalable k-means with Feature ranking (SKFR)

- The methods described before do not remove the uninformative features and keeps model uninterpretable.
- The authors proposed Sparse k-means with feature ranking (SKFR) to eliminate redundant features for clustering and keep interpretability intact.
  - It ranks features based by ordering d_l for each feature *l* with higher rank for feature with high d_l
  - Top s features are selected, rest features for centroid is set to zero

$$d_{jl} = \sum_{i \in C_j} (x_{il} - 0)^2 - \sum_{i \in C_j} (x_{il} - \mu_{jl})^2 = |C_j| \mu_{jl}^2$$

SKFR1 AND SKFR2





- SKFR1 (the global version) and SKFR2 (the local version). SKFR2 allows for the set of relevant features to vary across clusters. The informative components are updated within each cluster.

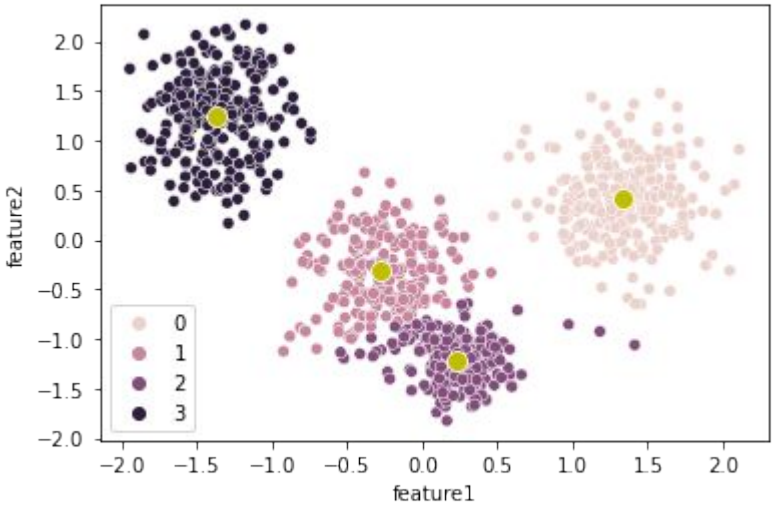
EXTENSIONS TO SKFR

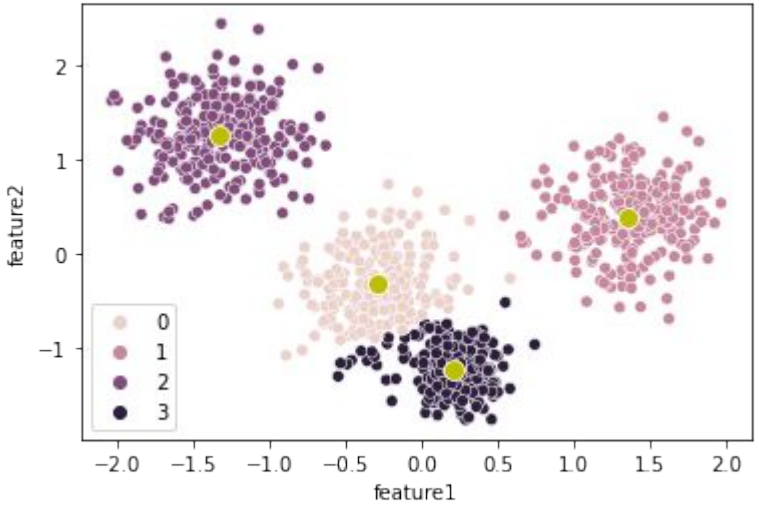
- SKFR works well not only with noisy, high dimensional data but also when the data contains **missing values** and **outliers**.
  - K-pod method along with SKFR can be used for imputing missing values, which makes best current guess of its value in each feature.
- 1-norm distance measure can be used instead of 2-norm distance for trimming outliers, which replaces the previous within-cluster means with within-cluster medians.
- Choice of sparsity can be determined by minimizing gap-statistic

$$\mathrm{Gap}(s) = \log O(\boldsymbol{X}, s) - \frac{1}{B}\sum_{b=1}^{B} \log O(\boldsymbol{X}_b, s)$$

$$O(\boldsymbol{X}, s) = \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2^2 - \sum_{j=1}^{k} \sum_{i \in C_j} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_2^2$$

SKFR IN ACTION





- Despite the different cluster colors, it can be seen that the two algorithms have very similar performance with some differences in performance visible in the outliers

Computational Complexity

- Cluster centers determine the rank of each of the features.
  - Therefore there is no need to go through the entire dataset again to calculate d_l when s = p.
- The algorithm can be shown to be the same as Lloyd's algorithm.
  - Therefore, the SKFR algorithms enjoy the same computational complexity as Lloyd's algorithm, O(npk), as the sorting algorithm has a logarithmic computational cost.

References

[1] Z. Zhang, K. Lange, and J. Xu, “Simple and scalable sparse k-means clustering via feature ranking,” in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 10 148–10 160. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/735ddec196a9ca5745c05bec0eaa4bf9-Paper.pdf

[2] D. Pollard, “Strong consistency of k-means clustering,” The Annals of Statistics, vol. 9, no. 1, 1981.

[3] J. T. Chi, E. C. Chi, and R. G. Baraniuk, “k-pod: A method for k-means clustering of missing data,” The American Statistician, vol. 70, no. 1, p. 91–99, 2016.

[4] “Skfr-python,” https://github.com/aarunishsinha/SKFR-Python, [Online; accessed 8-December-2021].