Evaluating Multi-view Kernel Clustering Algorithms

STUDENTS / UNIVERSITIES

Çağrı Eser — METU

Cazibe Kavalcı — Bilkent University

SUPERVISOR(S)

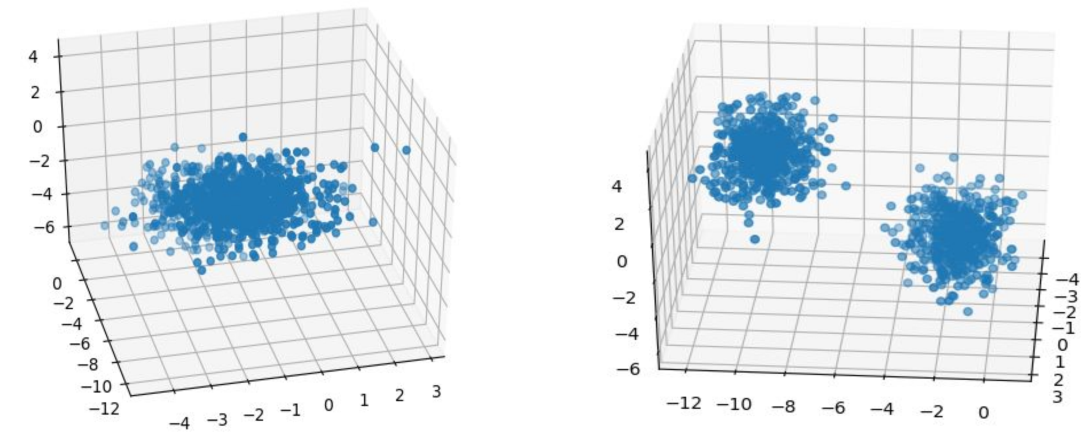
Öznur Taştan

Sabancı Üniversitesi

PURE — PROGRAM FOR UNDERGRADUATE RESEARCH

MULTI-VIEW KERNEL CLUSTERING

Clustering algorithms are used to understand the structure of unlabeled data by assigning each data point into a specific group. It is used in many fields like medical fields or even astronomics, which generally include data with multiple features.



The figure above includes two views of such data. From the first view, separation of data points cannot be observed, although from the second view it is clear. Therefore, multi-view clustering algorithms are needed to cluster multi-view data properly. Kernel functions are used to compute similarity of data while maintaining efficiency.

In the project, multi-view kernel clustering algorithms are evaluated. This evaluation aims to detect which algorithm to use while clustering since algorithms react differently to datasets or kernels with different properties.

Objectives

- Generating synthetic data with different levels of noise, randomness or views and finding appropriate real data to cluster.
- Experimenting with multi-view kernel clustering algorithms in terms of extrinsic evaluation metrics such as entropy, NMI and ARI using these data sets with different kernels.
- Analyzing findings and report performances of algorithms at experiments with different kind of data sets.

Methods

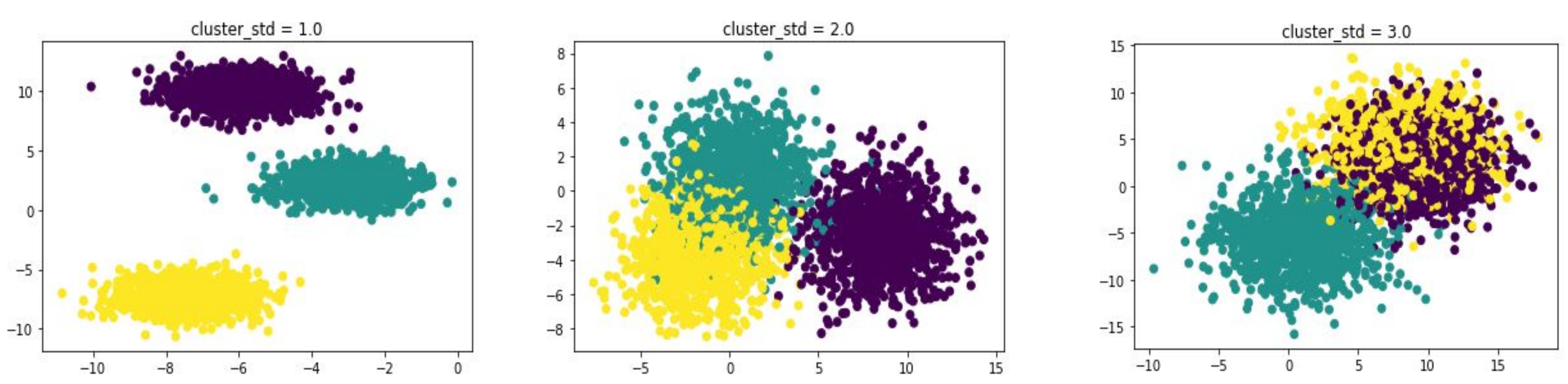
Algorithms used

Commonly used algorithms are selected and their structure, inputs and outputs are examined.

Algorithm	Description
SBKKM	Greedily chooses the best performing single kernel for k-means clustering.
AMKKM	Generates a new kernel by uniformly weighting all base kernels for clustering.
MKKM[1]	Alternatively performs kernel k-means and updates the kernel coefficients.
LMKKM[2]	Combines the base kernels by sample-adaptive weights.

Synthetic Data Generation

Data with different difficulty levels of clustering are generated with closer centers and increasing standard deviation to evaluate the algorithms.



Real Datasets Used

Dataset	Description	Samples
Flower17 dataset	A collection of images for 17 classes of flowers.	1360 images
BBC datasets	Collection of news articles in a pre-processed matrix format.	2225 documents
MNIST dataset	Handwritten digits (0-9) as 28x28 pixel images.	60.000 images

Experiments on Real and Synthetic Datasets

Evaluation metrics on real and synthetic datasets with different difficulty levels are as follows:

Adjusted Rand Index (ARI)

Algorithms	Flower17	BBC	MNIST	Easy	Medium	Hard
SBKKM	0.256	**0.822**	0.383	**1.000**	**0.832**	0.436
AMKKM	0.282	0.226	**0.385**	0.510	0.469	0.287
MKKM	0.297	0.742	0.361	**1.000**	**0.832**	**0.456**
LMKKM	**0.302**	0.464	0.364	**1.000**	0.752	0.410

Normalized Mutual Information (NMI)

Algorithms	Flower17	BBC	MNIST	Easy	Medium	Hard
SBKKM	0.443	**0.782**	0.476	**1.000**	**0.805**	0.397
AMKKM	0.464	0.276	0.477	0.592	0.529	0.347
MKKM	0.480	0.712	0.472	**1.000**	**0.805**	**0.413**
LMKKM	**0.481**	0.525	**0.480**	**1.000**	0.720	0.367

Entropy

Algorithms	Flower17	BBC	MNIST	Easy	Medium	Hard
SBKKM	4.409	**1.943**	3.493	**1.098**	1.309	1.744
AMKKM	4.332	2.547	3.489	1.544	1.589	1.788
MKKM	4.295	2.045	3.490	**1.098**	**1.309**	**1.729**
LMKKM	**4.284**	2.332	**3.461**	**1.098**	1.404	1.785

- The Flower17 dataset is evaluated with a set of precomputed similarity kernels.
- The BBC and MNIST datasets are evaluated with a family of RBF, polynomial and cosine similarity kernels.
- The Easy, Medium and Hard synthetic datasets are generated using the methods described in the previous section.

Conclusion

- Key observations for different kernel types are that:
  - Cosine kernels are better at clustering documents.
  - Polynomial kernels are better with images.
  - RBF kernels are more reliable for general clustering.
- General performances of the algorithms:
  - SBKKM is a greedy choice, but it provides a good baseline.
  - AMKKM's performance varies a lot with the kernels being used together.
  - MKKM and LMKKM performs highly similarly but MKKM clusters better while centers are closer and standard deviation is higher.

References

[1] Gönen, M., and Margolin, A. A. 2014. Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology. In NIPS, 1305–1313.

[2] Huang, H., Chuang, Y., and Chen, C. 2012. Multiple kernel fuzzy clustering. IEEE T. Fuzzy Systems 20(1):120–134.