How can I improve clustering performance.

- Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step.
- Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.
- Surprisingly for some cases, high clustering performance can be achieved by simply performing K-means clustering on the ICA components after PCA dimension reduction on the input data. However, the number of PCA and ICA signals/components needs to be limited to the number of unique classes.

What is the importance of clustering

In machine learning, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping unlabelled examples is called clustering. As the examples are unlabelled, clustering relies on unsupervised machine learning. If the examples are labelled, then clustering becomes classification.

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

- market segmentation
- social network analysis
- search result grouping
- medical imaging
- image segmentation
- anomaly detection

After clustering, each cluster is assigned a number called a cluster ID. Now, you can condense the entire feature set for an example into its cluster ID. Representing a complex example by a simple cluster ID makes clustering powerful. Extending the idea, clustering data can simplify large datasets. Machine learning systems can then use cluster IDs to simplify the processing of large datasets. Thus, clustering's output serves as feature data for downstream ML systems.