Data Mining report

Haoping Xiao, 888390
Haishan Wang, 913333

December 2020

Contents

1 Typical Gene Expression Data

1.1 Methods

1.1.1 Transformations and dimension reduction

We did not transform data or reduce dimensions of the data

1.1.2 Tested distance/similarity measures

We used Euclidean distances in all methods.

1.1.3 Tested clustering methods and parameters settings

We used two clustering methods on this data set: K-means clustering and Spectral clustering method. Parameters are set as follows:

1. K-Means clustering method: *sklearn.cluster.KMeans*
   - Parameters: *n_cluster=5*, other parameters are default.
   - The input data is original data set.
2. Spectral clustering method: *sklearn.cluster.SpectralClustering*.
   - Parameters: *n_cluster=5, n_neighbors=6, affinity='nearest_neighbors'*, other parameters are default.
   - The input data is original data set.

1.2 Results

1.2.1 Results Visualization

As table 1 shown, we have obtained precise results.

Methods	Distance/Similarity	Features reduction	NMI
K-Means	Euclidean distance	No	86.08%
Spectral	Euclidean distance	No	98.07%

Table 1: Comparison of K-Means and Spectral Clustering

1.2.2 Discussion about parameters selection

We set the number of clusters based on the prior information. When we reduced the dimension of data set from 7000 to 3 by PCA and then visualized it: Figure 1, we found the clusters compact and well-separated in Euclidean space. Therefore, we choose the Euclidean distance and K-means.

We optimize the number of neighbours of Spectral Clustering algorithm by

cross-validated grid-search over a parameter grid.

The validation of K-means algorithm is non-ideal, we infer the separating surface between clusters is spatially irregular. So we try the Spectral Clustering and construct affinity matrix by computing graphs of nearest neighbours, and obtained pleasant result.

2 Mass Spectrometry Data

2.1 Methods

2.1.1 Transformations and dimension reduction

Uniform manifold approximation and projection (UMAP) is a nonlinear dimensionality reduction technique. Principal component analysis (PCA) is commonly used for dimensionality reduction by projecting each data point while preserving as much of the data's variation as possible.

We choose this two type of transformations:

1. In the K-means Clustering, we normalized each feature of data, then performed UMAP on the data set to reduce the dimension of the data to 93.
2. In the Spectral Clustering, we performed PCA on the data set to reduce the dimension of the data to 694.

2.1.2 Tested distance/similarity measures

1. In the K-means clustering, We use the Euclidean distance.
2. In the Spectral clustering, We use the cosine similarity.

2.1.3 Tested clustering methods and parameters settings

We used two clustering methods on this data set: K-means clustering and Spectral clustering method. Parameters are set as follows:

1. K-Means clustering method: *sklearn.cluster.KMeans*,
   - Parameters: *n_clusters=3*, other parameters are default.
   - The algorithm input data is 93-dimensional, which is obtained by UMAP on normalized data.
2. Spectral clustering method: *sklearn.cluster.SpectralClustering*
   - Parameters: *n_clusters=3, affinity='precomputed_nearest_neighbors', n_neighbors=177*, other parameters are default.

- The algorithm input data is 694-dimensional, which is obtained by PCA on original data. We use the cosine similarity matrix of this input data to be the affinity matrix of the algorithm.

2.2 Results

2.2.1 Results Visualization

As table 2 shown, we have obtained precise results. For spectral clustering, the NMI score fluctuates between 97%-100% because of randomness of kmeans and umap.

Methods	Distance/Similarity	Feature reduction	NMI
K-Means	Euclidean distance	normalization & UMAP	98.22%
Spectral	Cosine Similarity	PCA	88.37%

Table 2: Comparison of K-Means and Spectral Clustering

2.2.2 Discussion about parameters selection

We set the number of clusters based on the prior information. When we visualized the median of 5000 features of data set: Figure 2, we found each feature falls in different ranges.

To eliminate dominant effect from features with large magnitudes, we transform features to the same range [0,1] using min-max scaling before the K-means method. For the same reason, we use cosine distance matrix as affinity matrix in the Spectral Clustering method.

We optimize the *number of components of UMAP* and *number of neighbours of Spectral Clustering algorithm*, by cross-validated grid-search over a parameter grid.

We obtain better clustering results from K-Means after UMAP on data set. Our explanation is UMAP keeps the essential features of mass-spectrometry data set.

3 Brief Instructions of Program

Use the following statement in Python code:

Step 1 Install package dependencies.
pip3 install -r requirements.txt

Step 2 Run *python3 cluster.py* in directory *./codes*

A Appendix Information Visualization

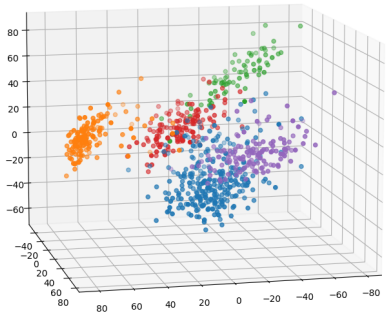

Figure 1: Visualize gene data after dimension reduction

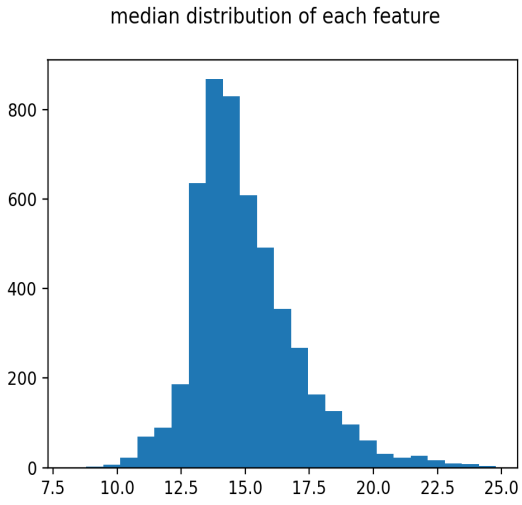

Figure 2: Visualize MS data median