perform_clustering

Overview

The perform_clustering.py script is designed to analyze ExpressionData files generated by BoolODE to cluster the data to predict the number of steady states that exist in the Boolean model. It currently supports k-means elbow, k-means silhouette, and DBSCAN.

Getting Started

If you have not installed the required packages mentioned in README.md, then you can do that with the command:

```
pip install -r requirements.txt
```

Tutorials with perform_clustering

DBSCAN Tutorial

Run the script as follows:

```
$ python perform_clustering.py -f absolute/path/of/ExpressionData.csv -d
```

If the file path you provided is correct, you will see the following message:

```
DBSCAN result:
Estimated number of clusters: 3
Estimated number of noise points: 19

DBSCAN analysis generated a DBSCAN_ClusterIDs.csv file.
```

The DBACAN_ClusterIDs.csv file should look like this:

```
,cl
E0_484,0
E1_1028,1
E2_2646,0
E3_872,1
E4_922,1
E5_669,1
E6_1070,2
E7_1771,0
E8_1628,-1
E9_198,2
E10_1762,2
E11_599,2
E12_433,0
E13_854,2
E14_1505,2
E15_2625,2
E16_1941,1
```

It is the same format as ClusterId files generated by BoolODE, where each cell is assigned to a cluster. Note that noise points identified by DBSCAN form their own cluster for visualization purposes.

Elbow of k-means Tutorial

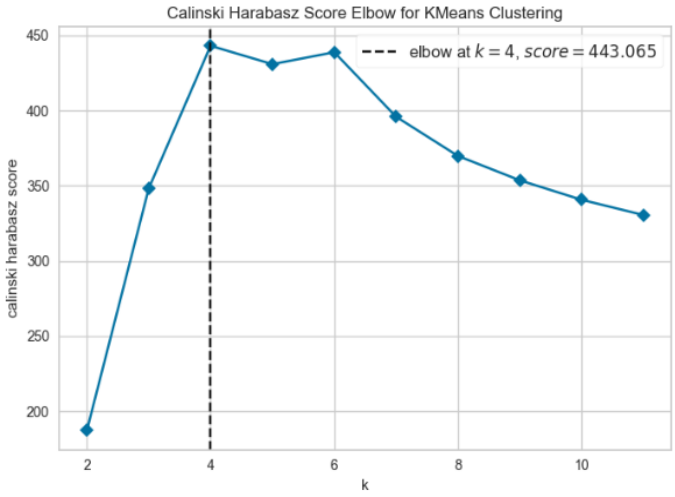
Run the script as follows:

```
$ python perform_clustering.py -f absolute/path/of/ExpressionData.csv -e
```

If the file path you provided is correct, you will see the following message:

```
Elbow analysis generated an elbow_visualization.png file.
```

The elbow_visualization.png file will look something like this:



Silhouette of k-means Tutorial

To specify the upper bound of clusters (in this example 3), run the script as follows:

```
$ python perform_clustering.py -f absolute/path/of/ExpressionData.csv -s -u 5
```

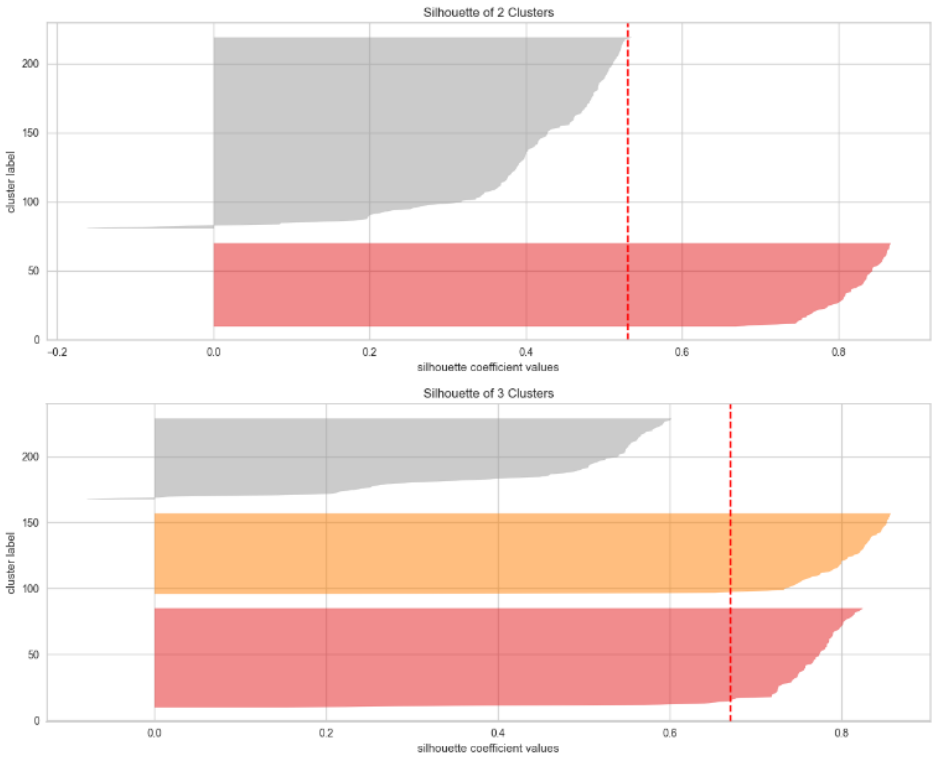
To run with the default upper bound number of clusters (11), run the script as follows:

```
$ python perform_clustering.py -f absolute/path/of/ExpressionData.csv -s
```

If the file path you provided is correct, you will see the following message:

```
Silhouette analyses generated a silhouette_visualization.png file.
```

The silhouette_visalization should look something like this (3 is the upper bound in this case):



How to interpret the silhouette_visualization.png file:

You would select the estimated value of k as the number corresponding to the plot with the least number of negative coefficient values and has the most uniformity in the thickness of the clusters. In this case, we would estimate 3 clusters.

Explanation: In the first plot, there is less uniformity for cluster thickness compared to the second plot, and there are more negative silhouette coefficient values.

Supplemental DBSCAN Explanation

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It is useful for handling the data in ExpressionData files because of how it handles noise. DBSCAN works by looking at a point and a specified number of its neighbors at a certain distance to determine if they are similar enough to be grouped together in a cluster. In the end, DBSCAN outputs the estimated number of clusters that it found from the data. DBSCAN is used to estimate the number of steady states from simulated gene expression data from an ExpressionData file