STAT 527 Nonparametric Regression and Classification Final Project

Joyce Lin, Student ID: 2225859
Department of Statistics, University of Washington
jlin99@uw.edu

June 4, 2023

Abstract

In this study, we investigated the application of mean-shift algorithm and random forests for unsupervised clustering. Additionally, we employed dimensionality reduction techniques, ISOMAP, to enhance clustering performance by capturing non-linear relationships and reducing noise. The results demonstrated clustering outcomes with distinct values observed in the clusters. The mean-shift algorithm showed stability in assigning data points to clusters, while the random forests approach exhibited room for improvement in terms of stability. However, after revealing that the dataset consists of 10 hand-written digits, it became apparent that the mean-shift algorithm yielded poorer clustering results compared to random forests. The mean-shift algorithm produced 21 clusters, which significantly exceeded the expected 10 clusters. In conclusion, Our findings highlighted the effectiveness of dimensionality reduction in enhancing unsupervised data analysis and emphasize the importance of further advancements in the stability of random forests clustering.

1 Introduction

Clustering techniques play a crucial role in uncovering hidden patterns and structures within datasets. In this study, we aim to explore and analyze the dataset using nonparametric classification methods within the framework of unsupervised learning. Nonparametric classification techniques provide flexibility in capturing complex patterns in the data without making explicit assumptions about the underlying distribution. The absence of labelled data in unsupervised learning poses a unique challenge, requiring algorithms to autonomously identify meaningful structures or clusters within the data. By leveraging nonparametric methods, we seek to uncover latent patterns and identify distinct groups or clusters that can be interpreted and assigned meaningful interpretations.

2 Data Description

The dataset used for analysis consists of 12,000 rows and 64 columns. Each row corresponds to a single data point that will be clustered, while each column represents a specific feature. A line plot (Figure 1) shows decreasing variability and range as dimensionality increases. The first dimension has higher variance and a wider range (variance: 4.968, range: -4.235 to 8.812) compared to the last dimension (variance: 0.115, range: -1.349 to 1.494). This suggests higher variable importance in earlier dimensions. Therefore, preprocessing and exploring data before clustering is crucial, and it is suggested to focus on the variables in the first few dimensions when considering their importance or impact on the clustering analysis.

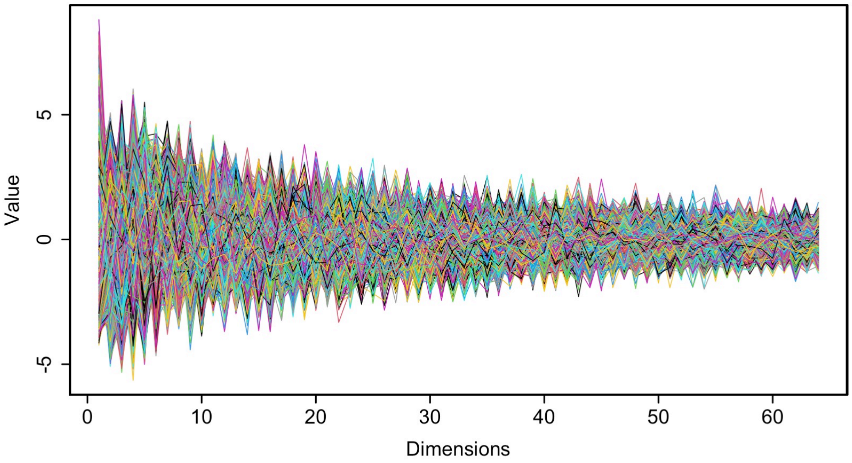

Figure 1: The line plot of the dataset, where x-axis represents the features and y-axis represents the values of the data points.

Upon unveiling the nature of the data, it was revealed that the dataset comprises 10 handwritten digits. In the upcoming sections, we will discuss clustering methods and results from the perspective of the unknown true number of clusters. The comparison between the clustering outcomes and the true results will be addressed in the discussion section.

3 Data Preprocessing

In our study with a dataset of 12,000 data points in 64 dimensions, dimensionality reduction is beneficial for clustering by preserving the underlying data structure. ISOMAP, which maintains geodesic distances between data points, is particularly useful for nonlinear manifold data, uncovering the underlying structure and preserving relationships between distant points [1]. Additionally, it retains cluster distinctiveness in the reduced space.

For ISOMAP, two important parameters are the number of neighbors and components. The number of neighbors determines the local structure preservation in the low-dimensional embedding [2]. By analyzing the scree plot of the reconstruction error (Figure 2), representing the error in reconstructing the dataset from the low-dimensional embedding, we select 15 neighbors based on the elbow point.

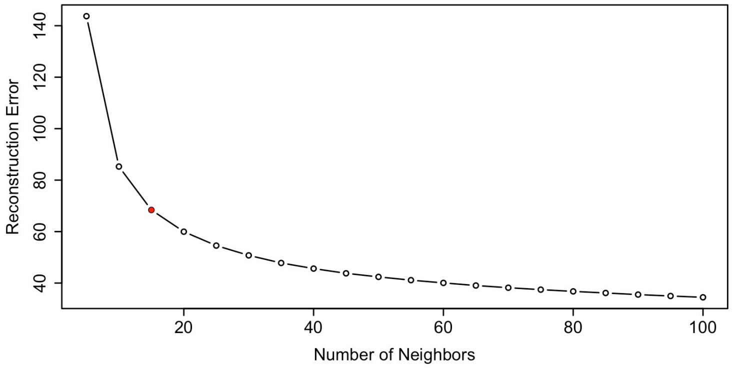

Figure 2: The scree plot of the reconstruction error in ISOMAP.

Furthermore, for visualization purposes, we reduce the dimensionality to 3 components, acknowledging the trade-off between interpretability and information loss. Although 3D may not capture all variance, it allows direct exploration and interpretation of the underlying structure.

4 Clustering Methods

In this study, we applied unsupervised clustering techniques to analyze the dataset. The specific clustering methods used were mean-shift and random forest, which will be introduced in Sections 4.1 and 4.2, respectively.

4.1 Mean-shift

The mean-shift algorithm is a clustering method that iteratively shifts the data points towards the local mean of their neighborhood, aiming to identify dense regions in the data. Since the data of our study is high dimensional, each data point is in $\mathbb{R}^d$, where $d = 64$, the dimensionality of our data. Furthermore, by exploring the data, we could observe that the data is smooth and symmetric, so the Guassian kernel is decided to use in this case. That is, we assume the Guassian kernel function $K(z) = \exp\left(\frac{-||z||^2/2}{\sqrt{2\pi}}\right)$, and we define the kernel density estimator function as

$$f_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h}),$$

where h is the bandwidth parameter. Also, the mean shift algorithm will be approximated through Newton's Method.

4.2 Random Forests and Hierarchical Clustering

Random forests, a popular machine learning algorithm, is employed in an unsupervised manner for clustering in our study. The algorithm constructs multiple decision trees by repeatedly sampling the training data with replacement and estimates classification error and variable importance using ”out-of-bag” (oob) data. Additionally, the algorithm computes proximities between pairs of cases, which represent the similarity or closeness between them.

The obtained proximity matrix is then subjected to hierarchical clustering. The chosen method for hierarchical clustering is the McQuitty method (WPGMA, Weighted Pair Group Method with Arithmetic Mean). In WPGMA, the distances between clusters are calculated as the weighted average distance between observations. This agglomerative approach merges similar data points or clusters, forming a hierarchical structure.

5 Result

In this section, we implemented the clustering methods using a specific strategy. Mean-shift clustering was applied to the raw data, while random forests were employed on the ISOMAP preprocessed data. We carefully selected key parameters for each method and assessed cluster stability, and visualization techniques were utilized to gain insights into the clustering outcomes.

5.1 Mean-shift with Raw Data

5.1.1 Clustering Strategy

For the mean-shift algorithm, we followed a specific strategy without performing data preprocessing. Our focus was on setting the key parameters for the mean-shift algorithm to optimize the clustering results. These parameters included the bandwidth (h), the number of neighbors, the number of iterations, and the scalar (ε) used for termination.

Firstly, we determined the bandwidth (h) of the kernel by employing maximum likelihood cross-validation. Using the `mkde.tune` function in `R`, we obtained an optimal bandwidth value of $h = 1.218$.

Next, we utilized the scree plot of the within-cluster sum of squares (WCSS) to select the appropriate number of neighbors. The WCSS measures the compactness of clusters by computing the sum of squared distances between each data point and its assigned cluster centroid. Lower WCSS values indicate more compact clusters. The scree plot of WCSS was shown in Figure 3. By evaluating the WCSS values for different numbers of neighbors, ranging from 1 to 30, we employed the elbow method and identified the number of neighbors as 21, which resulted in a lower WCSS.

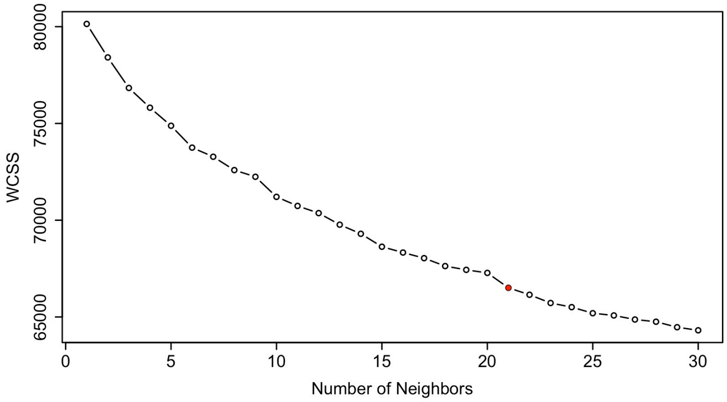

Figure 3: The scree plot of the WCSS for the selection of number of neighbors.

In addition, we set the number of iterations to 10 and $\varepsilon = 10^{-8}$ for individual query points in order to terminate the iteration process. These parameter choices were made to ensure convergence and achieve accurate clustering results. Finally, employing the mean-shift algorithm on the raw data using this clustering strategy resulted in the identification of 10 clusters.

5.1.2 Evaluation

To visualize the clustering results of the mean-shift algorithm, we generated a heatmap of the first 10 dimensions (Figure 4), which were determined to have higher importance in the data. It showcased distinct values among variables and clusters, indicating successful clustering. For instance, the first dimension of the 21st cluster had lower values compared to the 10th and 11th clusters, while the third dimension of the 10th cluster exhibited lower values compared to the 11th cluster. This visualization further confirmed the effectiveness of the clustering process.

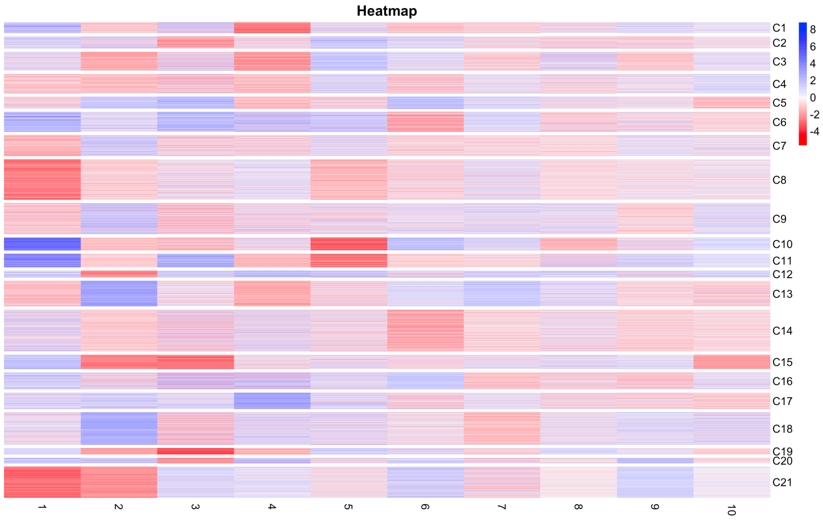

Figure 4: The heatmap of the clustering results obtained through the mean-shift algorithm, with variables on the x-axis and clusters on the y-axis.

To evaluate the stability of the mean-shift clustering results, we introduced noise to perturb the original data ($\mathcal{D}$) and generated new data ($\mathcal{D}'$). The Misclassification Error (ME) distance was then calculated to assess the stability. The noise was generated from a normal distribution, $\varepsilon \sim N(0,\ \sigma^2 I)$. To ensure that the noise remained much smaller than the diameter of the smallest cluster, we set σ to be 0.001 times the diameter of the smallest cluster. After obtaining the perturbed data, we re-applied the mean-shift algorithm and computed the ME distance, which resulted in $d_{ME} = 0.005$. This small distance suggests that the majority of data points remained assigned to the same clusters in both the original and perturbed datasets.

5.2 Random Forests and Hierarchical Clustering with ISOMAP

5.2.1 Clustering Strategy

Random forests are effective in capturing linear relationships between variables, but they may struggle to capture complex, non-linear relationships inherent in the data. To address this limitation, dimensionality reduction techniques such as ISOMAP can be employed to unveil and preserve non-linear relationships in a lower-dimensional space. By reducing the dimensions, noise can be filtered out, allowing the focus to be placed on the most relevant features. Consequently, we preprocessed the data using ISOMAP and reduced the dimensionality to three dimensions.

Regarding the parameters of the random forests algorithm, we considered the number of trees to grow and the number of variables randomly sampled as candidates at each split. To determine the number of trees, assuming the number of groups to be 15, we computed the WCSS and observed that the lowest value was obtained when using 1000 trees (Figure 5). Additionally, as the task in this study involves classification, it is recommended to set the number of variables randomly sampled at each split to $\sqrt{p}$, where p represents the total number of variables.

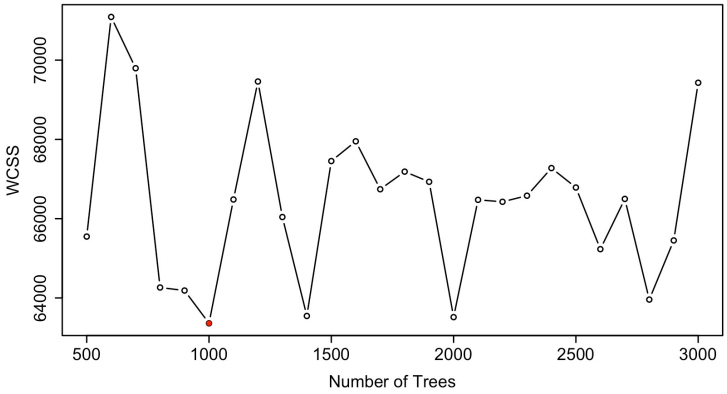

Figure 5: The scree plot of the WCSS for the selection of number of trees in the random forest.

For hierarchical clustering, the parameter that needs to be determined is the number of groups. By utilizing the scree plot of the WCSS, we employed the elbow method and identified 14 as the optimal number of groups (Figure 6). This decision was based on evaluating the WCSS values for different numbers of groups, ranging from 1 to 30.

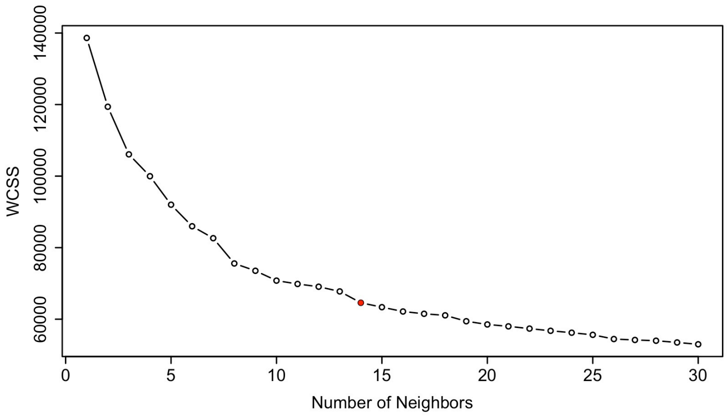

Figure 6: The scree plot of the WCSS for the selection of number of trees in the random forest.

5.2.2 Evaluation

To visualize the clustering results obtained from random forests, we utilized a 3D scatter plot to examine the clusters after reducing the dimensions to 3. As shown in Figure 7, although certain data points may exhibit close proximity within their respective clusters, the overall performance of the random forests algorithm in accurately assigning data points to their appropriate clusters is evident.

In addition, as shown in Figure 8, the heatmap displayed distinct values among different clusters and variables, which means that we get successful clustering results. This visualization provides additional evidence supporting the efficacy of the clustering approach.

We also conducted a comparative analysis of the clustering results obtained from random forests using the raw data and the dimension-reduced data. To evaluate their performance, we employed a heatmap visualization (Figure 9). The clustering results based on the raw data exhibited imbalanced clusters with less distinct outcomes, highlighting the limitations of using the raw data for clustering purposes.

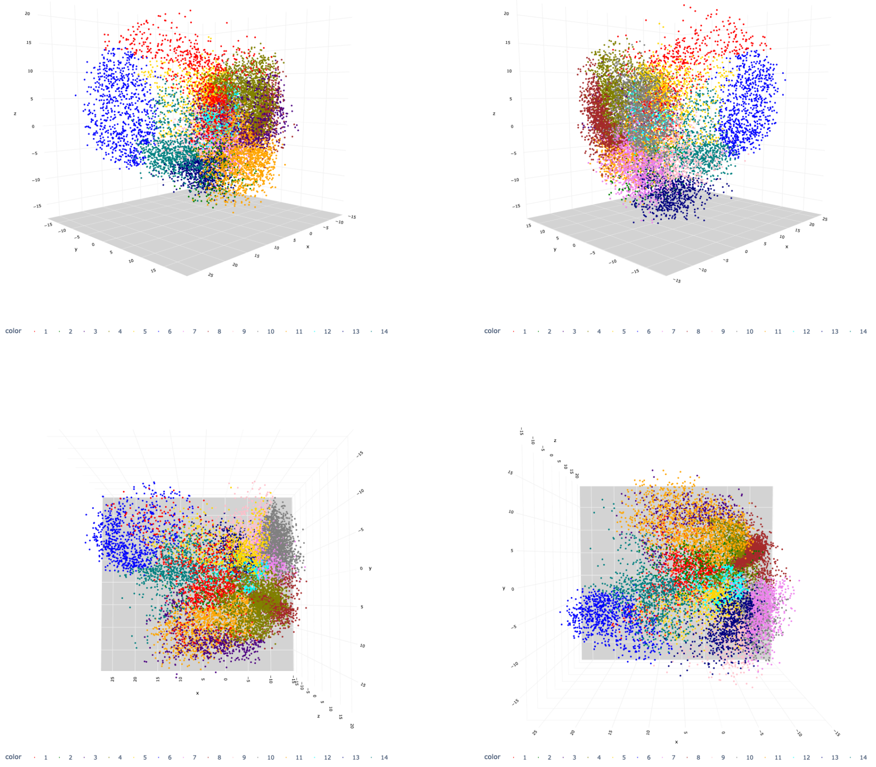

Figure 7: The 3D plots in different angles of the dimension-reduced data along with the clustering results obtained from random forests.

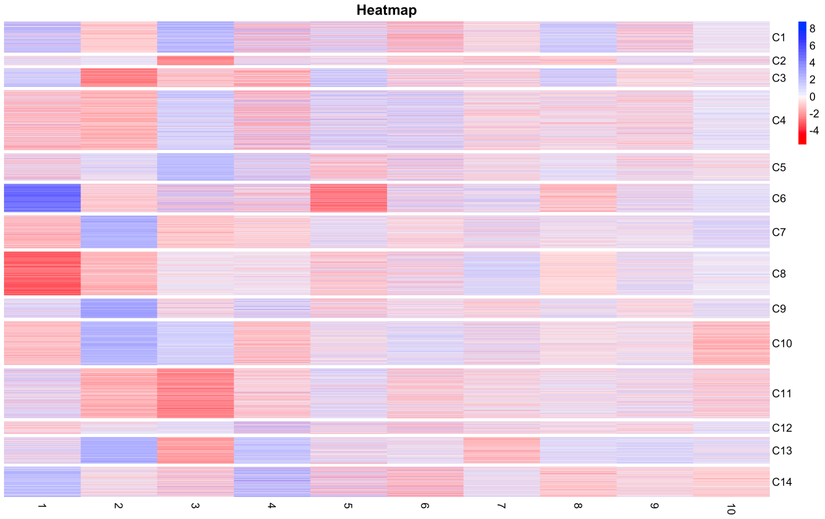

Figure 8: The heatmap of the clustering results obtained through the random forests based the dimension-reduced data, with variables on the x-axis and clusters on the y-axis.

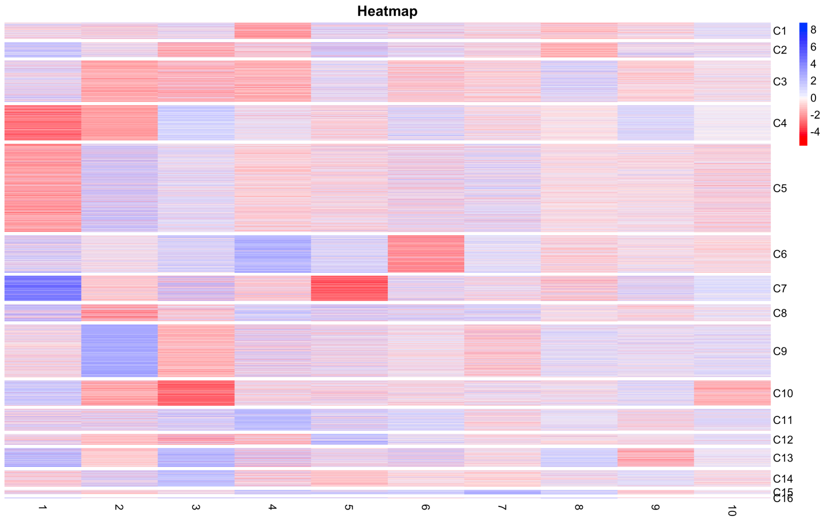

Figure 9: The heatmap of the clustering results obtained through the random forests based the dimension-reduced data, with variables on the x-axis and clusters on the y-axis.

Furthermore, to assess the stability of the random forests clustering results, we introduced noise to the dimension-reduced data by sampling from a normal distribution, $\varepsilon \sim N(0,\ \sigma^2 I)$, where σ is 0.001 times the diameter of the smallest cluster. Subsequently, we re-applied the random forests algorithm to the perturbed data and calculated the Variation of Information (VI) distance and Normalized Mutual Information (NMI) as evaluation metrics, yielding values of 3.258 and 0.574, respectively. The NMI value of 0.574 represents the level of similarity between two clusterings. A higher NMI value indicates a greater agreement between the clusterings. In this case, an NMI of 0.574 suggests a moderate level of agreement, indicating that the cluster assignments have some shared information but are not highly consistent and reliable.

6 Discussion

In this study, we applied the mean-shift algorithm and random forests to analyze a dataset and identify meaningful clusters within the raw data and the reduced-dimensional data by ISOMAP.

The use of ISOMAP for dimensionality reduction proved beneficial in uncovering non-linear relationships and filtering out noise. This resulted in providing valuable insights into the underlying patterns and relationships within the data. The visualization of the reduced-dimensional data allowed for a comprehensive understanding of the clusters and facilitated informed decision-making for subsequent analyses.

The mean-shift algorithm demonstrated its effectiveness in identifying clusters in the raw data. The heatmap visualization (Figure 4) clearly showed distinct values between different clusters and variables, indicating successful clustering outcomes. Additionally, the high stability of the mean-shift clustering method was evident from the low Misclassification Error (ME) distance, suggesting that the majority of data points remained assigned to the same clusters even after introducing perturbations to the data.

Random forests and hierarchical clustering offered an alternative to unsupervised clustering. Random forests captured complex relationships and important variables, while hierarchical clustering revealed underlying structure using computed proximities. The heatmap (Figure 8) showed distinct values in clusters, indicating successful clustering. However, stability evaluation uncovered inconsistencies due to densely clustered data points in the dimension-reduced data, challenging stability when adding noise. Enhancements are necessary for reliable and stable random forests clustering.

Upon uncovering the true nature of the data as hand-written digits, it became evident that the actual number of clusters should be 10. This revelation highlights the importance of preprocessing and exploring the data before engaging in clustering analysis. However, it should be noted that the mean-shift algorithm produced 21 clusters, which deviated significantly from the expected 10 clusters. This discrepancy may have resulted from an overly objective approach in selecting the number of neighbors based on the within-cluster sum of squares scree plot. Despite the low ME observed in the mean-shift method, reassessing the parameter selection methodology is crucial due to the substantial deviation from the expected outcome. One potential approach is stability selection, which involves introducing perturbations to create random subsets and assessing instability to determine the appropriate number of clusters. Incorporating stability selection enhances reliability and provides a more accurate representation of the underlying structure in the hand-written digits data.

Future research can focus on exploring different methods for parameter selection and refining the clustering process to align the results more closely with the true number of clusters.

References

[1] Silva, V., & Tenenbaum, J. (2002). Global versus local methods in nonlinear dimensionality reduction. Advances in neural information processing systems, 15.

[2] Samko, O., Marshall, A. D., & Rosin, P. L. (2006). Selection of the optimal parameter value for the Isomap algorithm. Pattern Recognition Letters, 27(9), 968-979.

[3] Von Luxburg, U. (2010). Clustering stability: an overview. Foundations and Trends in Machine Learning, 2(3), 235-274.