Aalto University

CS-E4650 Methods of Data mining

Project work

Danh Phan - 673987
Khanh *Linh Nguyen - 814762

Table of Contents

1. Methods

At the beginning, we tried to cluster the data sets (both gene data and ms data) without pre-processing them. As predicted, no clustering methods yielded a good result, except Spectral Clustering for Gene Data. As the data sets contain a vast number of features, the aim of this process is to project the data to two-dimension space. Since the feature names are cryptic and we do not know anything about the domain knowledge, selecting two features having highest correlation values is feasible, but is not really meaningful in this case. Hence, dimension reduction techniques were used, which are: t-SNE and UMAP. One of the main differences between these two is that UMAP guarantees the data global structure better than t-SNE (Oskolkov, N). In this exercise, we want to find out the most appropriate combination of a dimension reduction technique with a clustering method based on the Normalized Mutual Information (NMI).

t-SNE: By using an appropriate similarity metric, t-SNE creates a probability distribution that similar instances will have high probabilities and dissimilar instances will have lower ones. Afterwards, t-SNE projects the instances on low-dimensional space while minimizing KL divergence between the distributions.

UMAP: By assuming that the data has uniform distribution in Riemannian manifold, the Riemannian metric is approximately constant, and manifold is locally connected; it models a manifold with a fuzzy topological structure. UMAP projects data on a low-dimensional space in a way that it has the closet possible fuzzy topological structure to the original (McInnes, L, Healy, J).

For clustering process, a variety of clustering algorithms were utilized including Agglomerative, Birch, Gaussian Mixture Model and K-Mean. The NMI score obtained by each method will be reported and discussed.

2. Results

2.1 Clarification on parameters

After testing different metric parameters for both data sets, it was confirmed that “cityblock” (Manhattan) yielded the best result when using t-SNE technique, and “Canberra” (in “Miscellaneous spatial metrics” type) gave the most outstanding result for UMAP method.

Since the parameter method was set to ‘exact’ to get the lowest nearest-neighbor errors, the only one need to be examined is “perplexity” when calling t-SNE. By looping through different values, one-dimension array was constructed for each clustering methods and the highest NMI score was found.

UMAP receives min_dist and n_neighbors as parameters. By looping through different combinations, two-dimension array was constructed for each clustering methods and the highest NMI score was found.

2.2 Gene Data

- ***Data exploration***

The gene dataset has 795 samples with 7002 features. Data contains only numerical values and there are no invalid data fields. As shown in the Sum Square Error Elbow from figure 1, the line indicates that 5 is the optimal cluster number since the speed slows down there.

- ***Data transformation and clustering***

t-SNE

The NMI scores of all clustering methods combined with dimensional reduction using t-SNE are presented in table 1 below. The best NMI results were observed in Agglomerative, Birch and Gaussian Mixture methods when utilizing t-SNE technique. The NMI values reached 0.986. The perplexity parameter was 15 for Agglomerative and Birch and 35 for Gaussian Mixture, n_clusters equaled to 5 with threshold was put at 0.1. The pcolor plot of three clustering methods are presented in figure 2. Spectral clustering has the lowest NMI value of 0.961. Figure 3 shows the t-SNE technique with Spectral Clustering pcolor plot.

	Agglomerative C	Birch	Gaussian M	K-Means	Spectral C
NMI	0.986	0.986	0.986	0.968	0.961

TABLE 1. Gene data - T-SNE NMI comparison

UMAP

The NMI scores of all clustering methods combined with dimensional reduction using UMAP are presented in table 2 below. The results show that UMAP techniques work extremely well with all clustering processes which gave a significant high NMI score of over 0.989. Figure 4 are pcolor plots of all methods. Since all methods have the best score, K-Means configuration will be discussed quickly. The optimal configuration for UMAP is min_dist = 0.3 and n_neighbors = 10, KMeans uses n_clusters equal to 5, init='k-means++'.

	Agglomerative C	Birch	Gaussian M	K-Means	Spectral C
NMI	0.989	0.989	0.989	0.989	0.989

TABLE 2. Gene data – UMAP NMI comparison

- ***Conclusion***

The non-optimal data in low-dimension using t-SNE technique can be seen from figure 5. Three algorithms including Agglomerative, Birch and Gaussian Mixture with t-SNE technique gave the highest NMI results, their clustering visualizations are displayed in figure 6. The reason behind can be explained according to the overview of clusterings (Scikit-learn), Agglomerative and Birch algorithms especially suits well for large dataset with many features and clusters. Gaussian Mixture, which is a versatile method for different data scale, also works well this dataset.

The non-optimal data in low-dimension using UMAP technique can be seen from figure 7. With UMAP, all methods performed clustering process very well as indicated from their NMI values. The observed NMI score even higher than all values recorded with t-SNE technique. This indicates that this dimensional reduction technique is the more effective on the gene data set than t-SNE since now the data is more separated and does not have any special shapes, such as spiral shape.

2.3 MS Data

- ***Data exploration***

The MS dataset has 694 samples with 5002 features. Data contains only numerical values and there are no invalid data fields. As shown in the Sum Square Error Elbow from figure 8, the line indicates that 3 is the optimal cluster number since the speed slows down there.

- ***Data transformation and clustering***

t-SNE

The highest NMI value, which is 0.9398, was found at Spectral Clustering. In this case, perplexity for t-SNE was set to 10 and the n_clusters parameter for Spectral Clustering was 3. T-SNE with Spectral Clustering pcolor plot can be found in figure 9. Birch technique, which has the lowest NMI value, is demonstrated with pclor plot in figure 10. In this case, perplexity for t-SNE is set to 15 and Birch has n_clusters is equal to 3 and threshold is equal to 0.1. Additionally, after trying threshold from range 0.1 to 1, 0.1 gave the best result with the value of 0.7509. Below is the table which compares different clustering techniques.

	Agglomerative C	Birch	Gaussian M	K-Means	Spectral C
NMI	0.8958	0.7509	0.9377	0.9141	0.9398

TABLE 3. MS data – t-SNE NMI comparison

UMAP

The pcolor plot of Gaussian Mixture technique, which has the best NMI value of 0.9182, is shown in figure 11. The optimal configuration for UMAP is min_dist = 1 and n_neighbors = 15. Table 4 below compares different clustering techniques.

	Agglomerative C	Birch	Gaussian M	K-Means	Spectral C
NMI	0.9054	0.9073	0.9182	0.9146	0.9046

TABLE 4. MS data – UMAP NMI comparison

- ***Conclusion***

Using t-SNE, from the data representation in 2D in figure 12, it can be speculated that “Spectral Clustering” and “Gaussian Mixture” techniques will be the best candidates for clustering in this case. The reasons are that the clusters are mostly separated with very few overlapped points; those clusters do not have special shapes, such as spiral shape; and the number of clusters are small. They performed quite well with the NMI values are all over 0.93. Birch method, as expected, did not do well in the clustering task since its use case was shown in the figure 14.

With UMAP, since there are several points closer to different groups than it class as shown in figure 15, even the Gaussian Mixture already did its best as illustrated in figure 16, the result is still lower than the Spectral Clustering using t-SNE.

Brief instructions for running the program

umap-learn and scikit-learn-extra are used and they need to be installed.

Functions generate_tsne_embedding() and generate_umap_nmi_score_matrix () are time consuming (10 and 70 minutes, respectively) since they generate score matrices to find out the best configurations. Therefore, sections 1.3.1.1, 1.3.2.1, 2.3.1.1 and 2.3.2.1 should not be run. For reproducibility, random_state was set to 42 when possible.

References

McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018.

Oskolkov, N. “tSNE vs.UMAP: Global Structure”. [Online] https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17. [Accessed 25.11.2020].

Wikipedia, "Curse of dimensionality," [Online]. Available: https://en.wikipedia.org/wiki/Curse_of_dimensionality. [Accessed 22 11 2020].

Scikit Learn, "Clustering," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html. [Accessed 24 11 2020].

Appendix

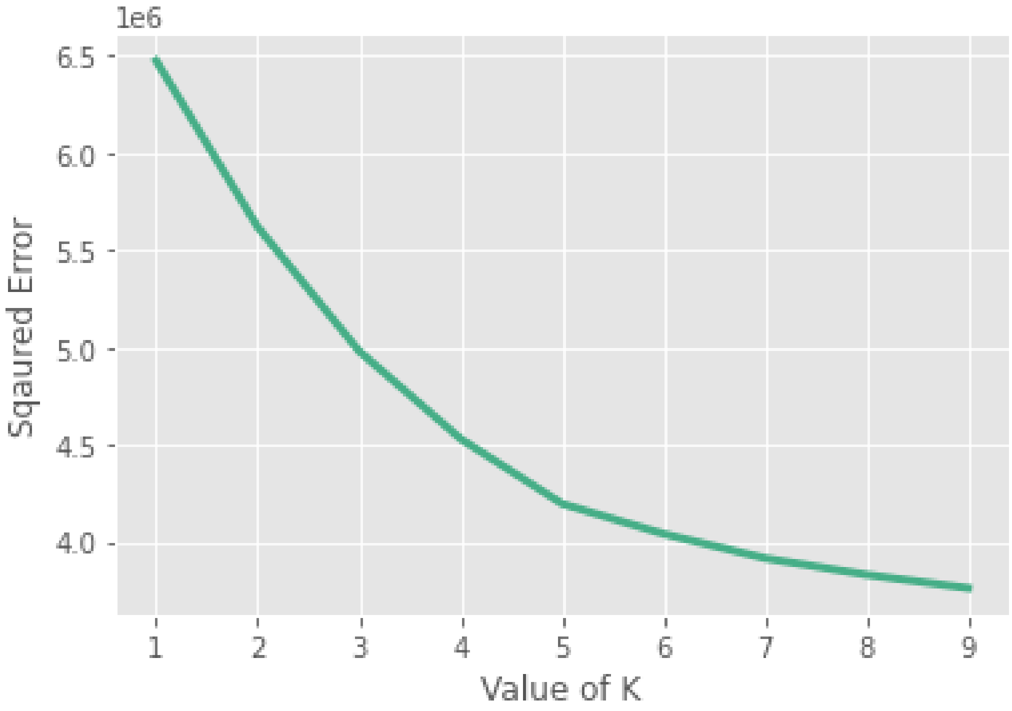
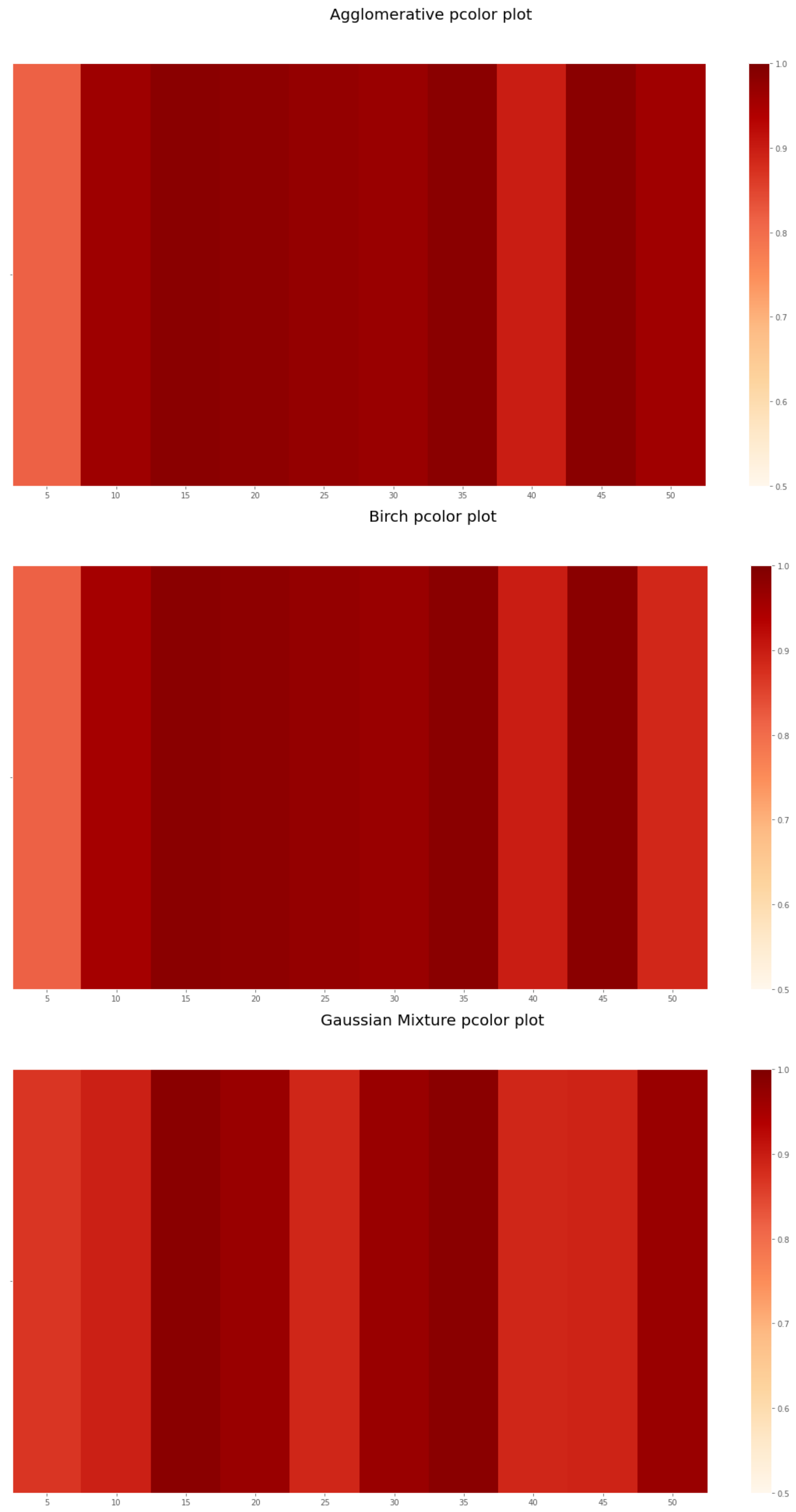


FIGURE 1. SSE elbow for gene data



Agglomerative pcolor plot

Birch pcolor plot

Gaussian Mixture pcolor plot

FIGURE 2. Gene Data - T-SNE cityblock Agglomerative, Birch and Gaussian Mixture Clustering pcolor plots

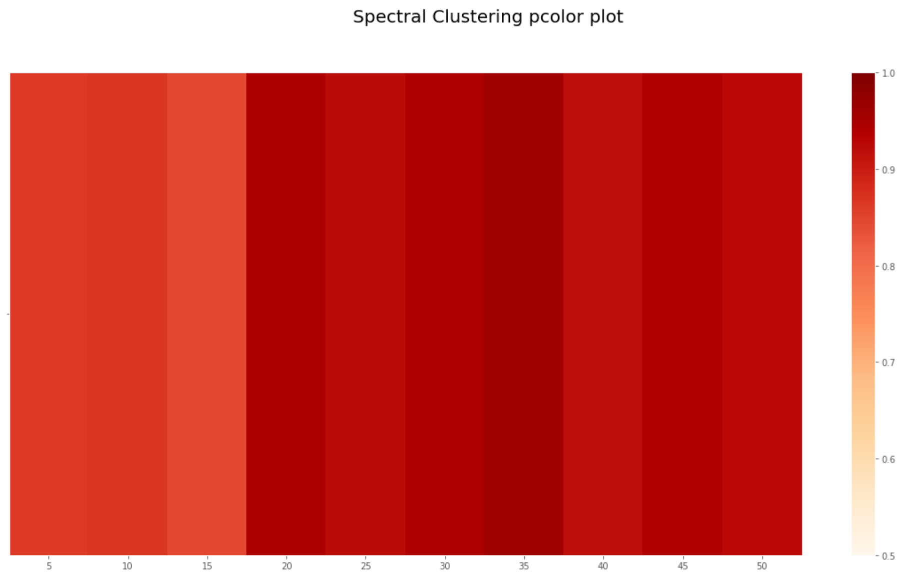


FIGURE 3. Gene Data - T-SNE cityblock Spectral Clustering pcolor plot

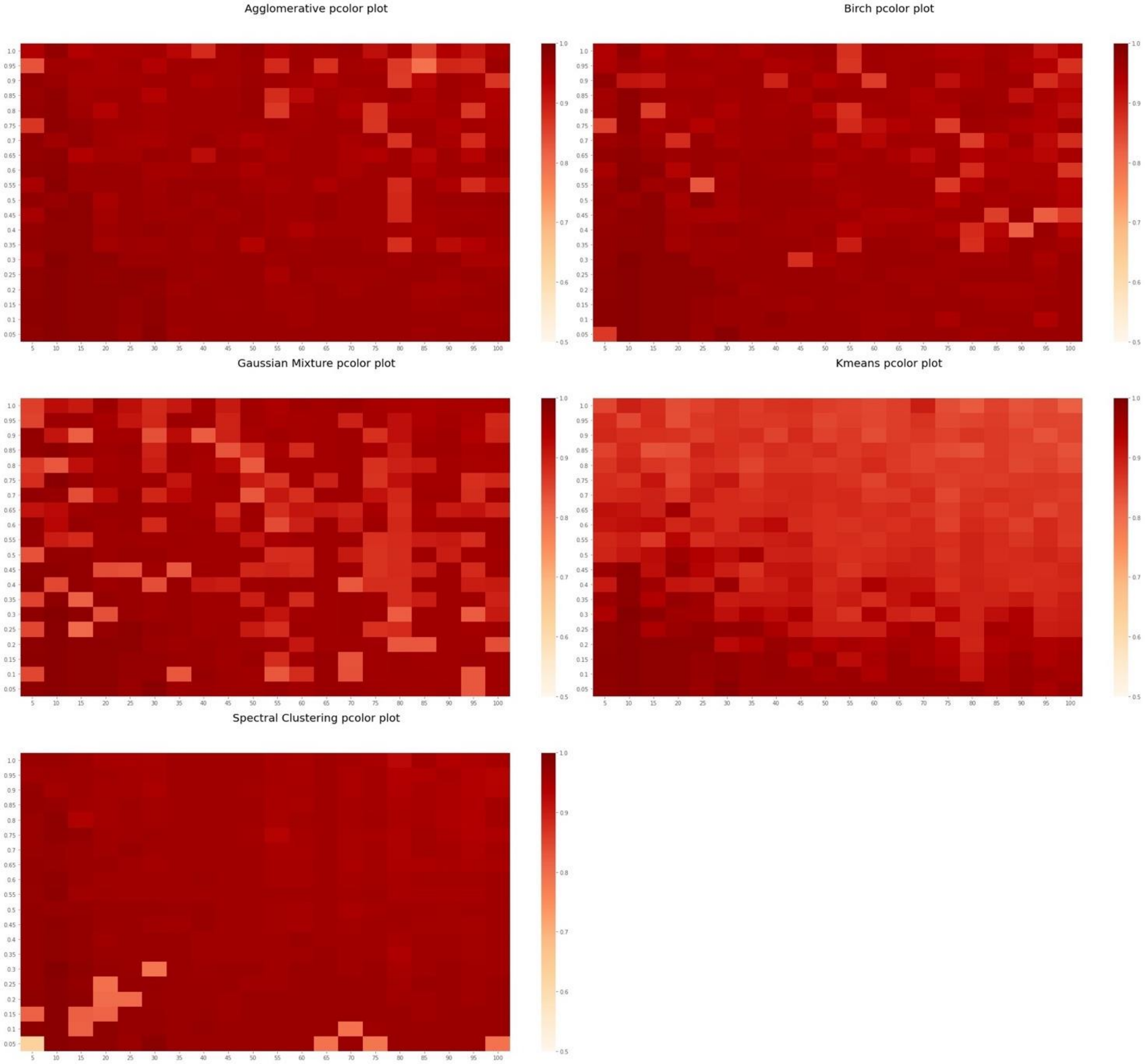


FIGURE 4 . Gene Data – UMAP Canberra pcolor plots

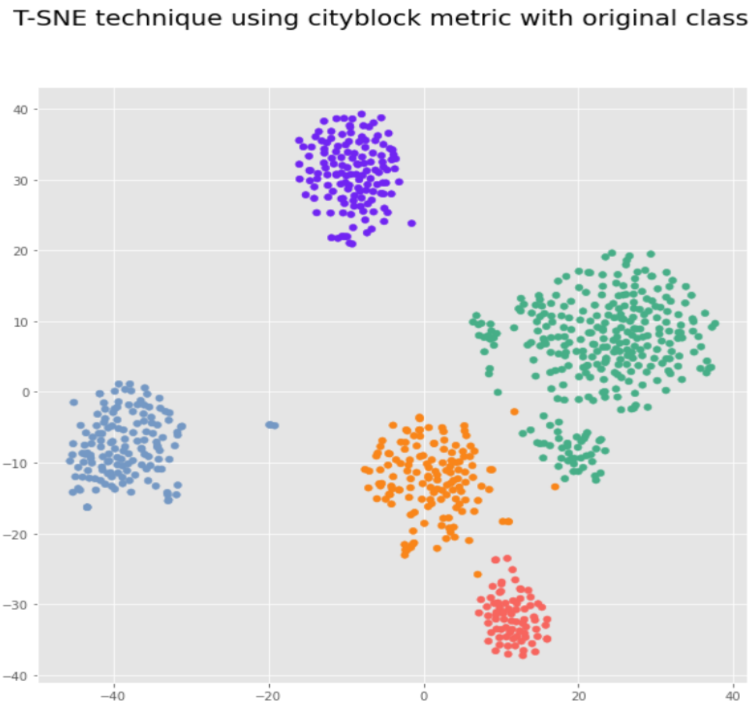
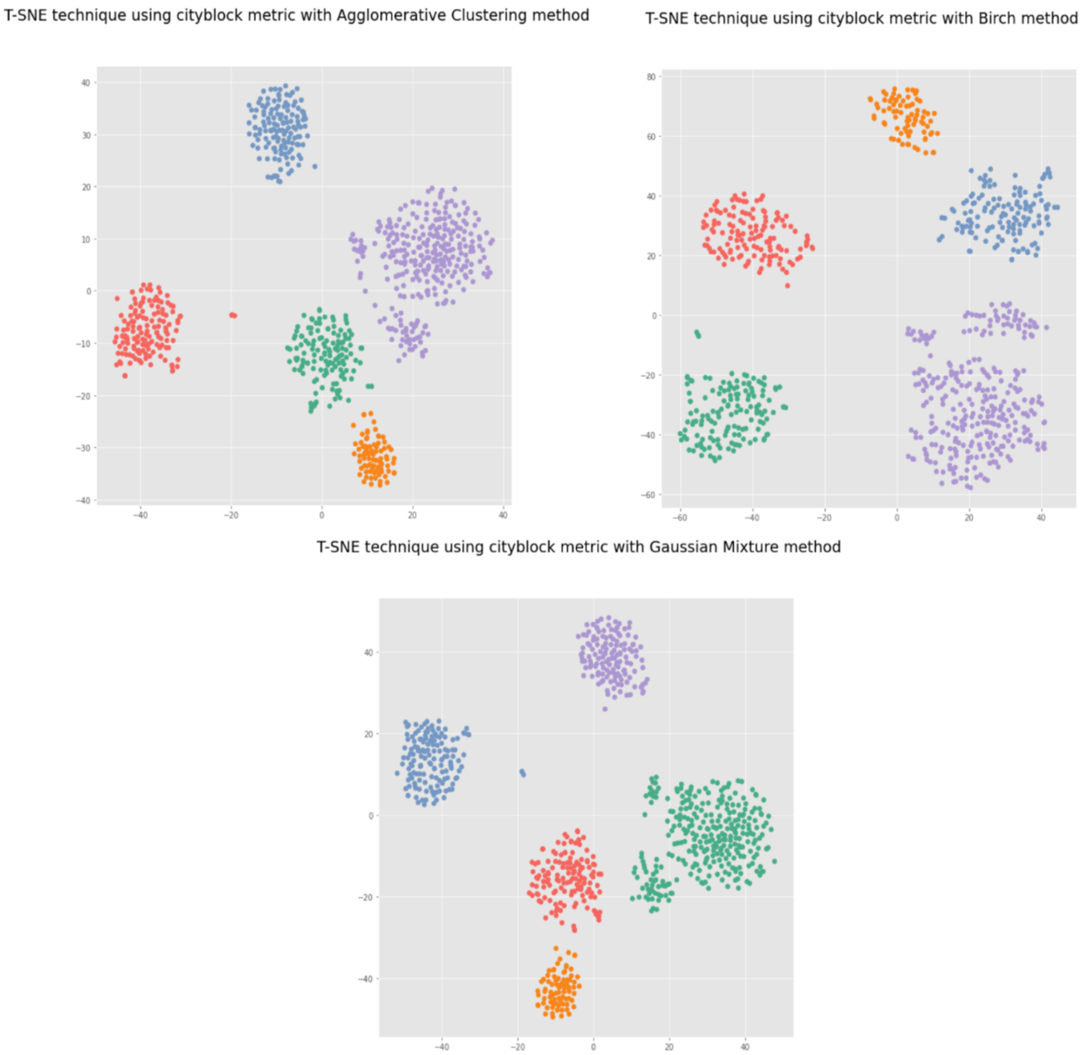

T-SNE technique using cityblock metric with original class

FIGURE 5. Gene data - T-SNE technique using cityblock metric with original class


T-SNE technique using cityblock metric with Agglomerative Clustering method

T-SNE technique using cityblock metric with Birch method

T-SNE technique using cityblock metric with Gaussian Mixture method

FIGURE 6. Gene data - T-SNE technique using ‘cityblock’ metric with Agglomerative, Birch and Gaussian Mixture clustering

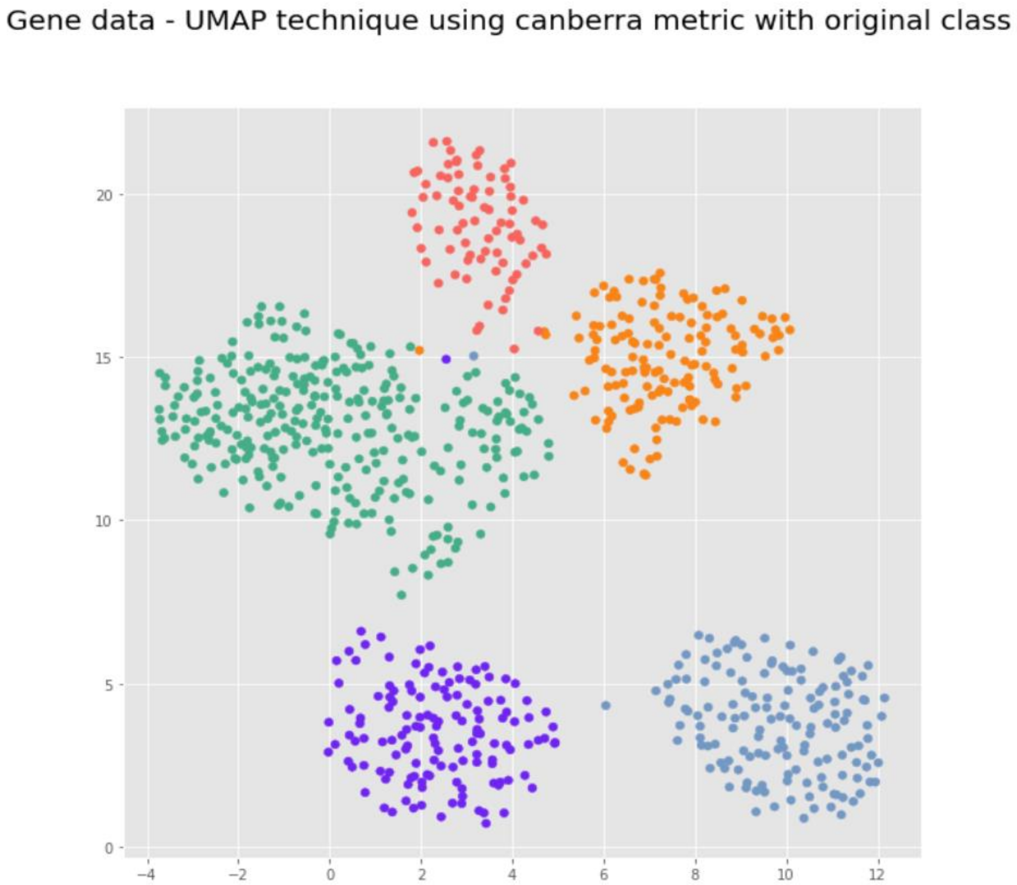


FIGURE 7. Gene data - UMAP technique using Canberra metric with original class

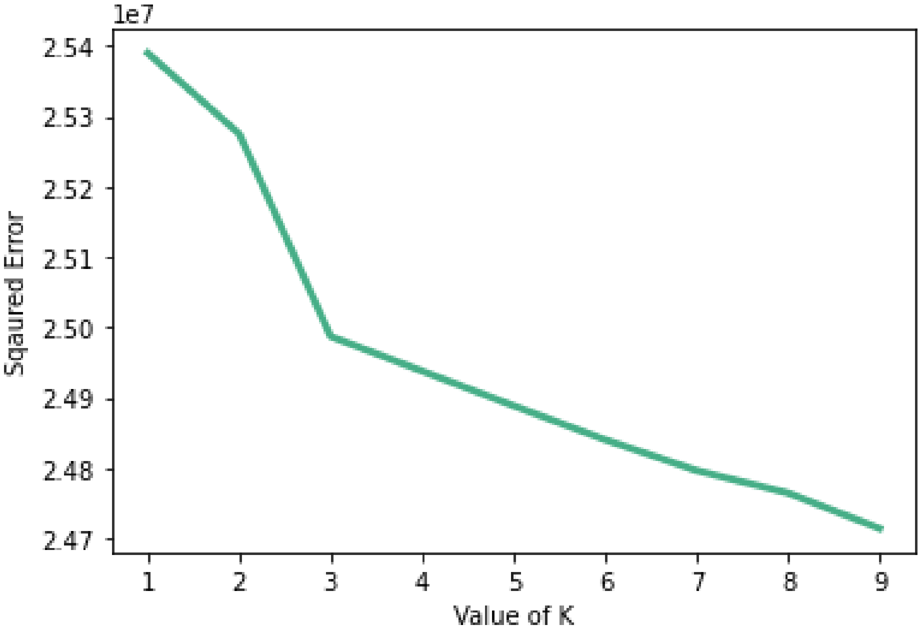


FIGURE 8. SSE elbow for MS data

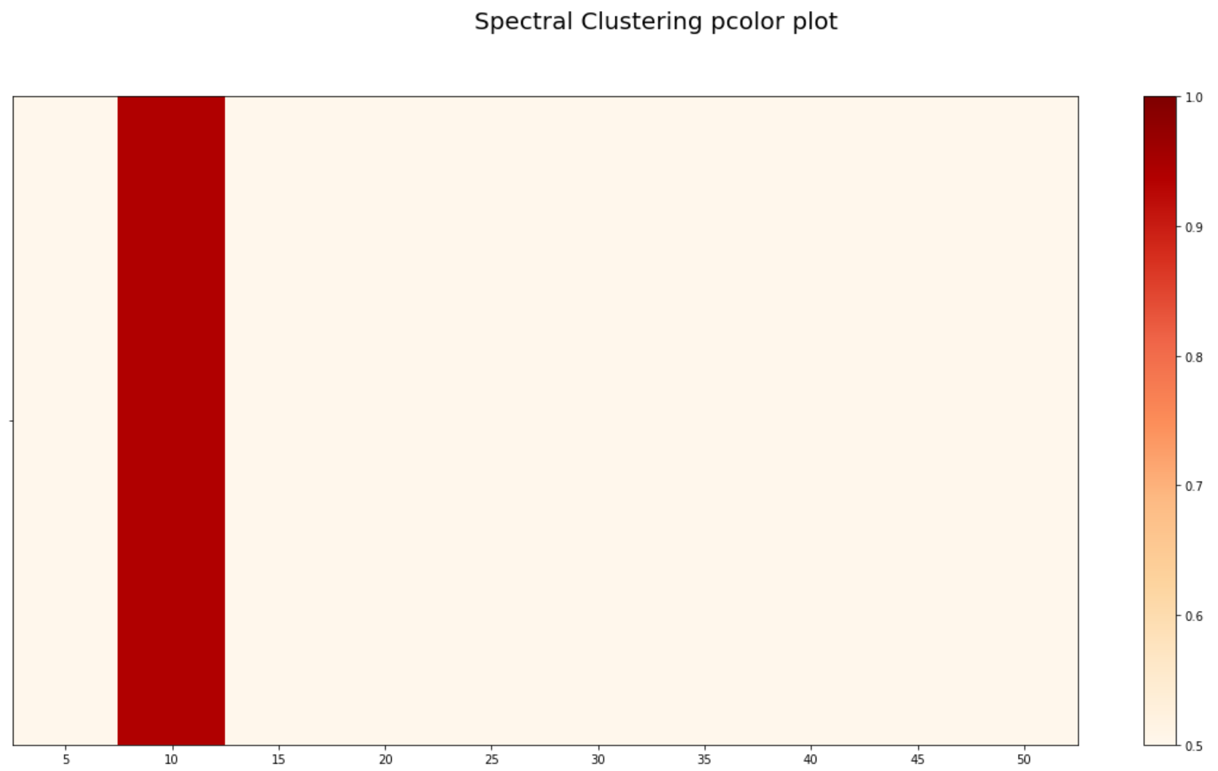


FIGURE 9. MS Data - T-SNE cityblock Spectral Clustering pcolor plot

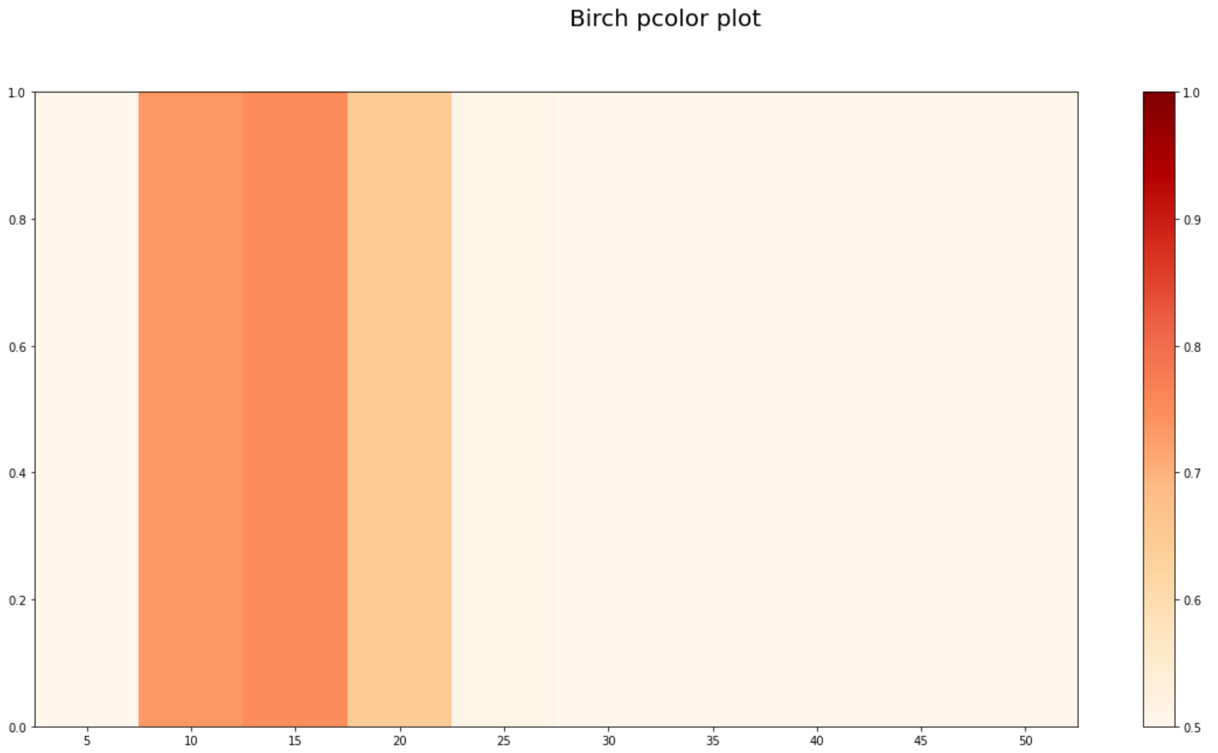


FIGURE 10. MS Data - T-SNE cityblock Birch pcolor plot.

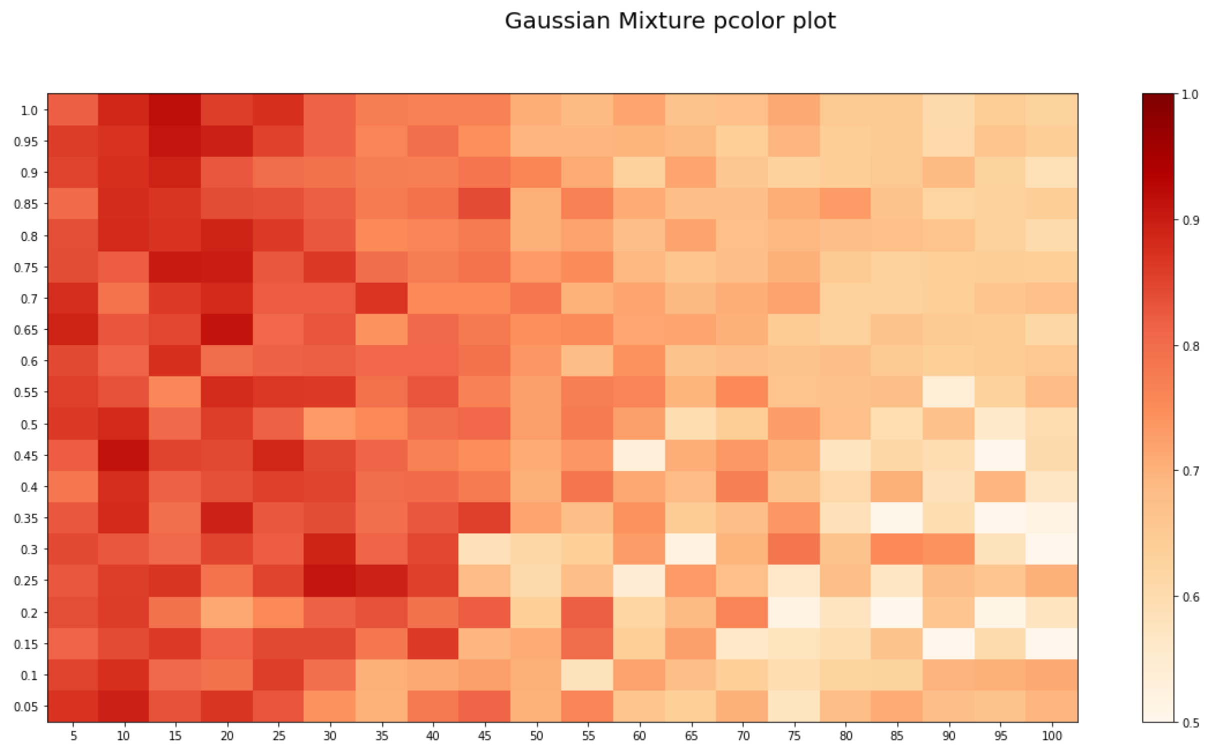


FIGURE 11. MS Data - UMAP Canberra Gaussian Mixture pcolor plot.

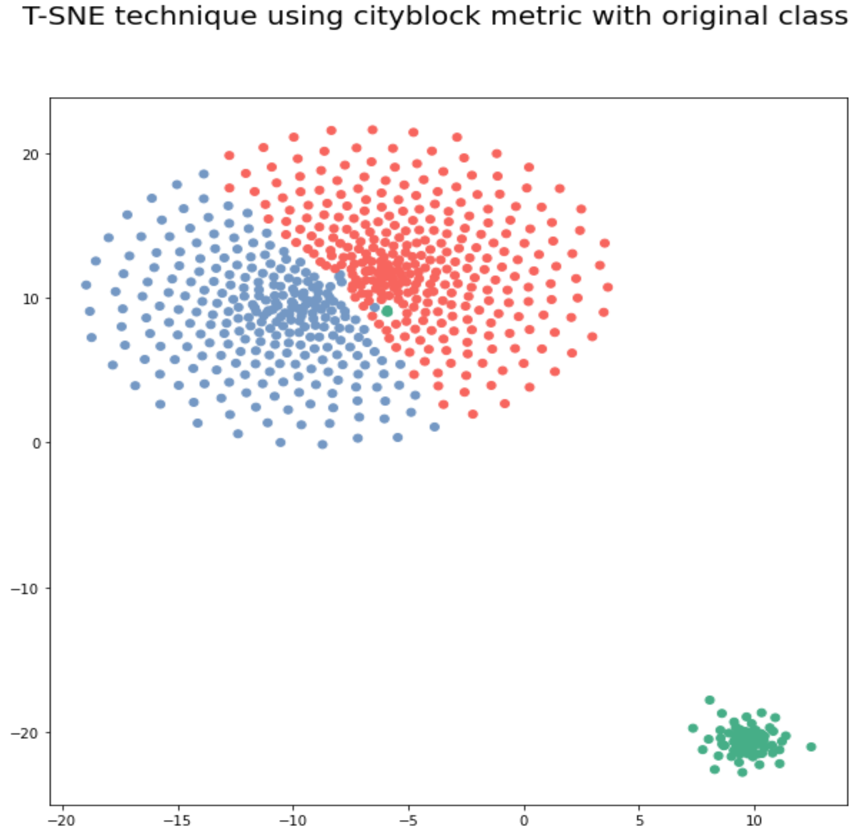


FIGURE 12. MS Data - T-SNE technique using ‘cityblock’ metric with original class

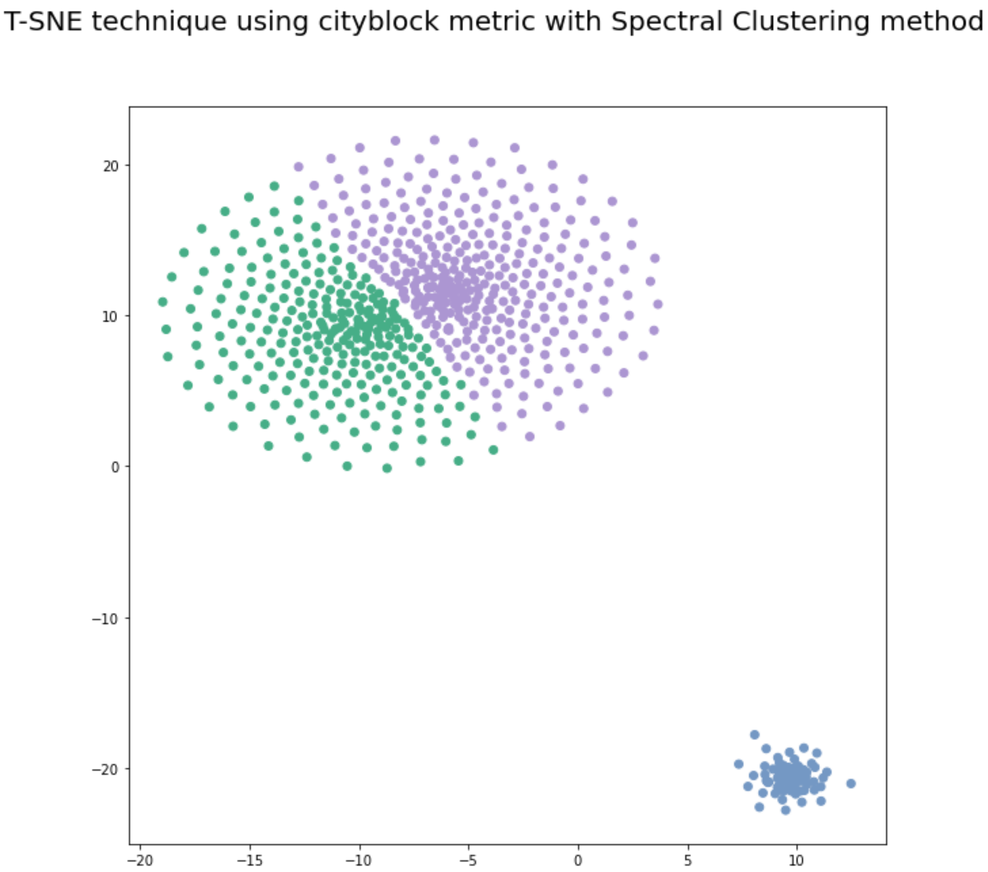


FIGURE 13. MS Data - T-SNE technique using cityblock metric with Spectral Clustering method

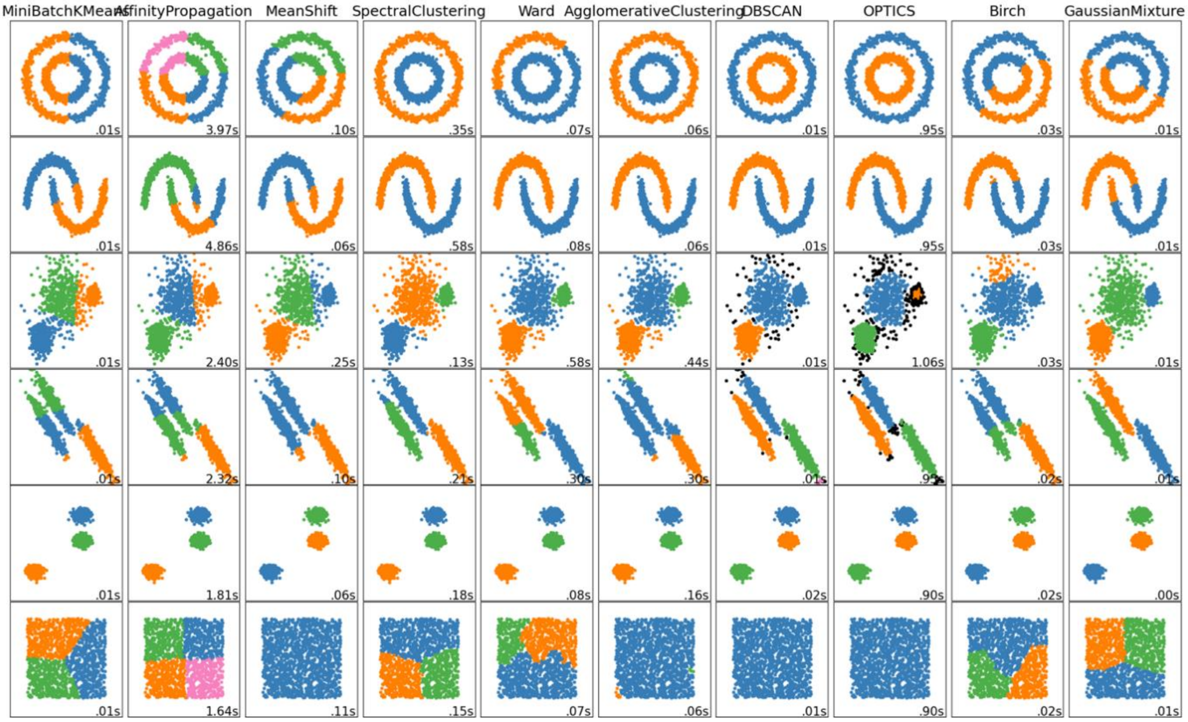


FIGURE 14. Overview of clustering methods [2]

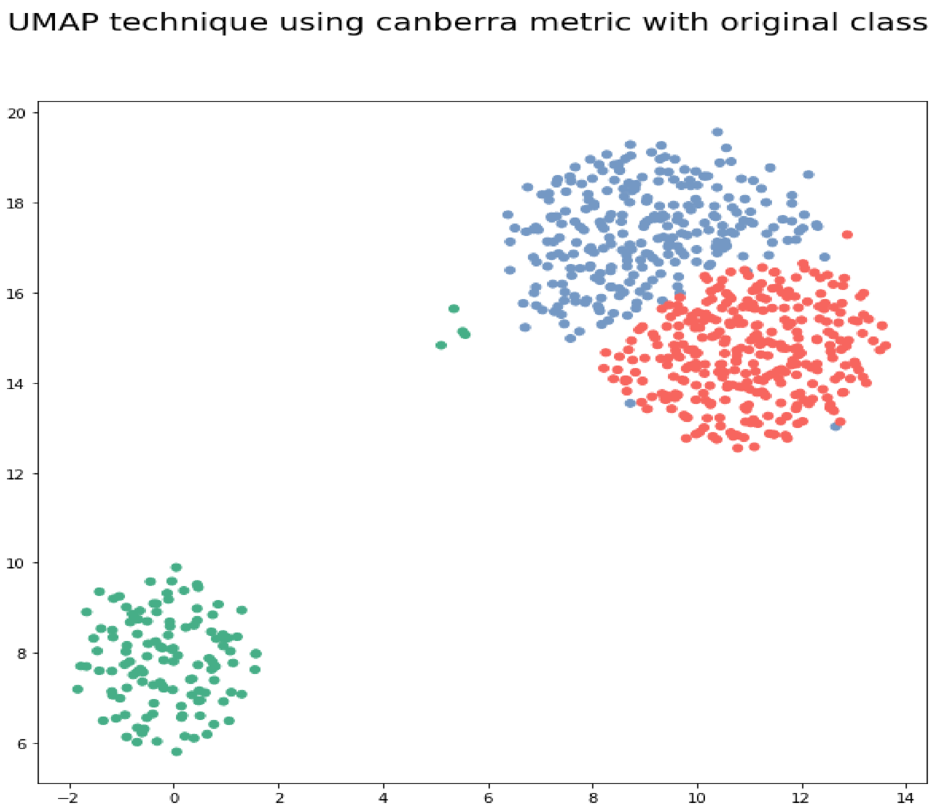


FIGURE 15. MS Data - UMAP technique using Canberra metric with original class

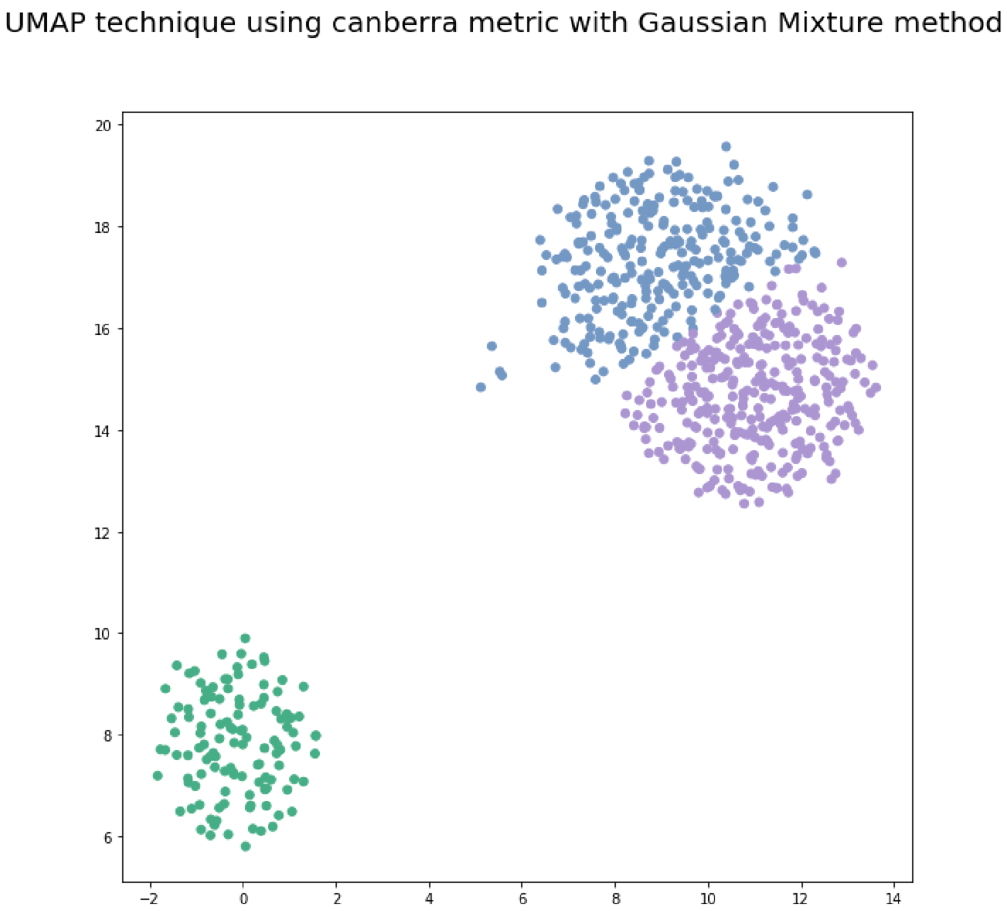


FIGURE 16. MS Data - UMAP technique using Canberra metric with Gaussian Mixture method