Competency Assessment Assignments

Juin 2, 2022

Hao ZHANG

Task 1 _a

Choose of approaches

- Because the original cancer datasets do not contain annotations, the effect of joint dimensionality reduction approaches on factors with clinical annotations or biological annotations does not need to be considered when processing such datasets. Just need to consider the effect of factors on survival.
- In general, MCIA, RGCCA, and JIVE achieved the best performances, finding factors significantly associated with survival in seven out of ten cancer types.

Task 1 _a

Choose of datasets

- Between the five multi-omics cancer datasets, I'd live to choose BIC (dataset of breast), LIHC (dataset of the liver) and SKCM (dataset of melanoma).
  Because MCIA performed the best for melanoma; JIVE performed the best in liver cancer ; RGCCA performed the best in breast cancer.

Task 1 _a

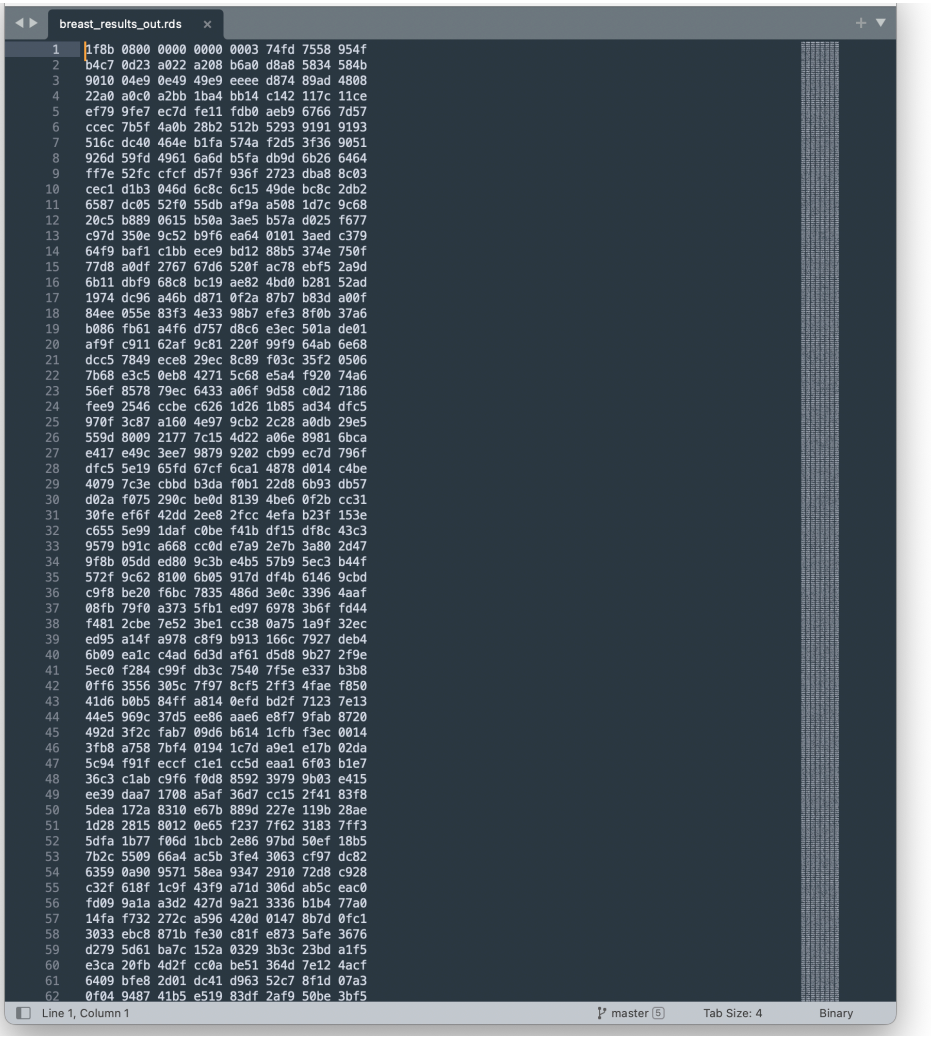
Description of datasets

- The Cancer Genome Atlas (TCGA), has profiled thousands of tumor samples for multiple molecular assays, including mRNA, microRNAs, DNA methylation, and proteomics. The cancer TCGA data were downloaded from http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. These data are composed of three omics layers: gene expression, DNA methylation, and miRNA expression. The number of samples ranges from 170 for Acute Myeloid Leukemia (AML) to 621 for Breast cancer.

Task 1 _a

Results sample

The jDR factorization will decompose the P omics matrices into a product of a single factor matrix F and multiple weight/projection matrices Ai, i=1...P.



Task 1 _b

Choose of approaches

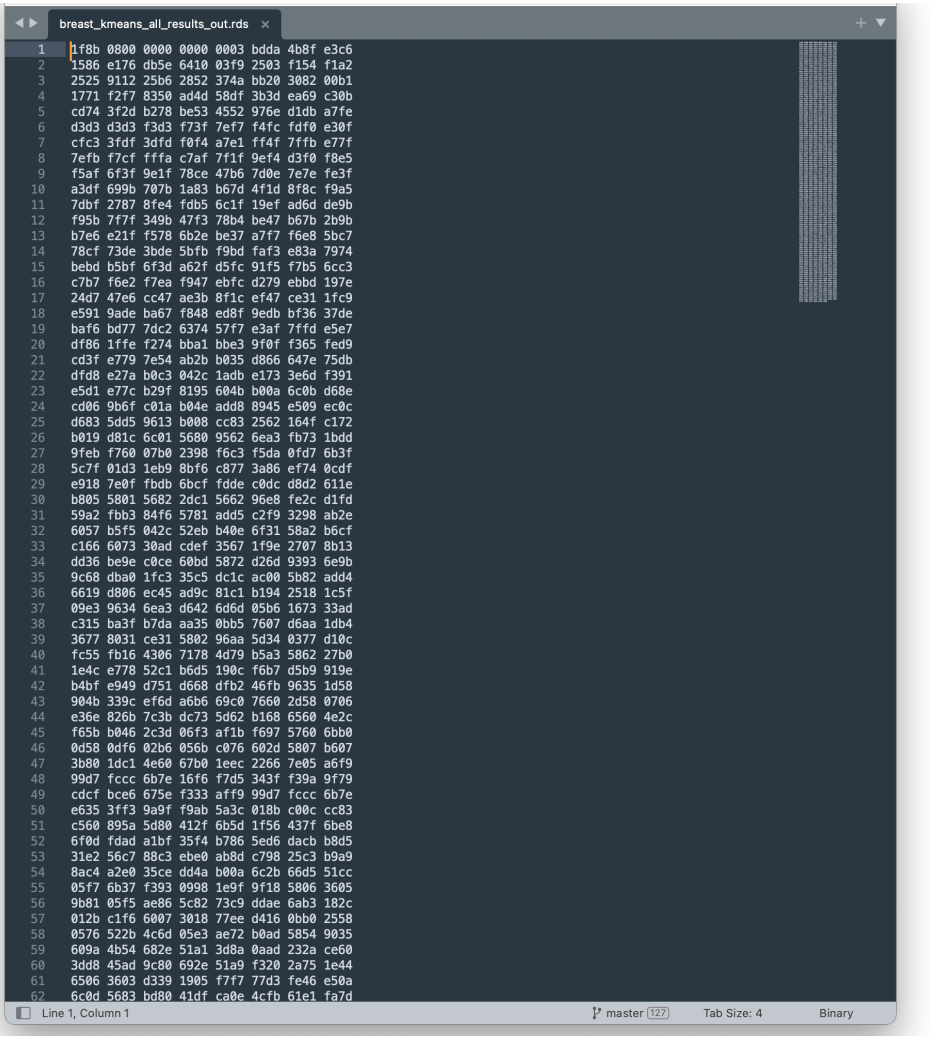
- No algorithm consistently outperformed all others in either differential survival or enriched clinical parameters. With respect to survival, MCCA had the best prognostic value, while MultiNMF was second and LRACluster third.

- However, limited by the installation of the package and the program provided, I can only choose K-means, Spectral and SNF for clustering.

Task 1 _b (1)

Results sample

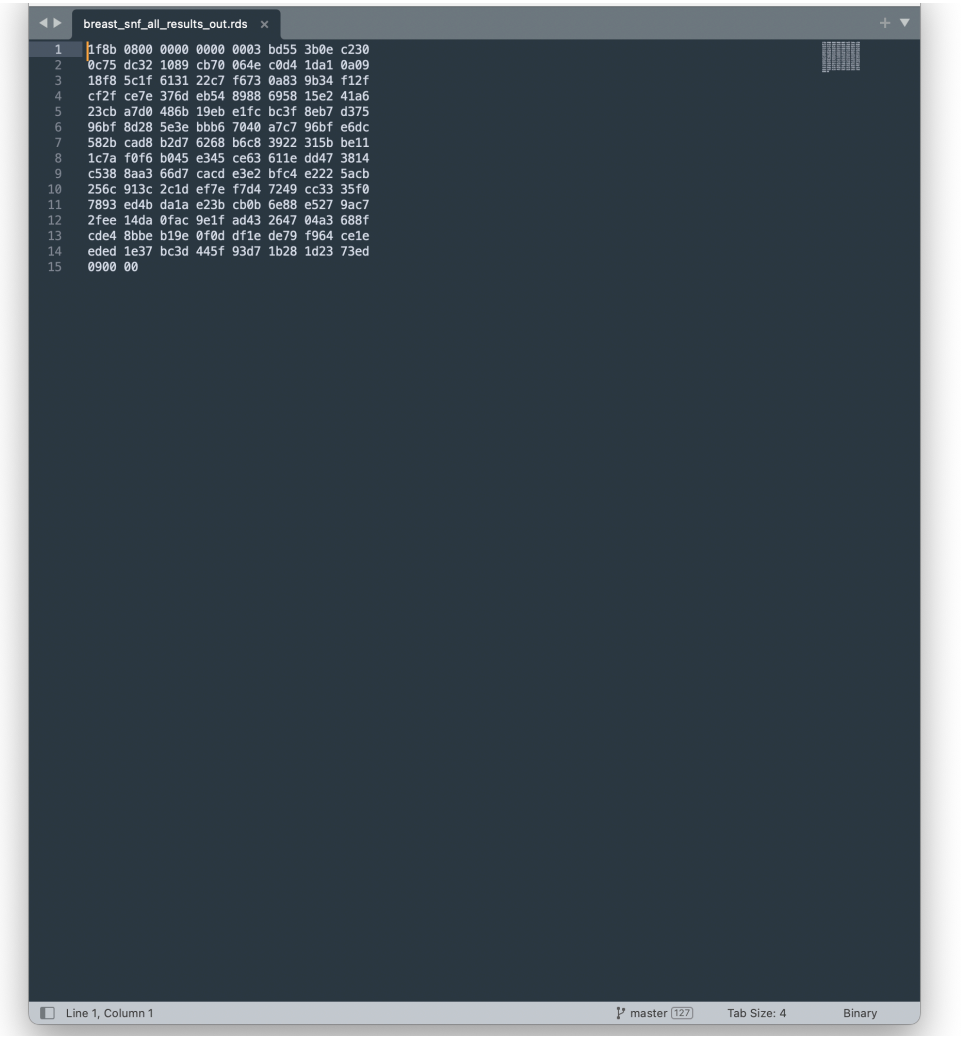
Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters.



Task 1 _b (2)

Results sample

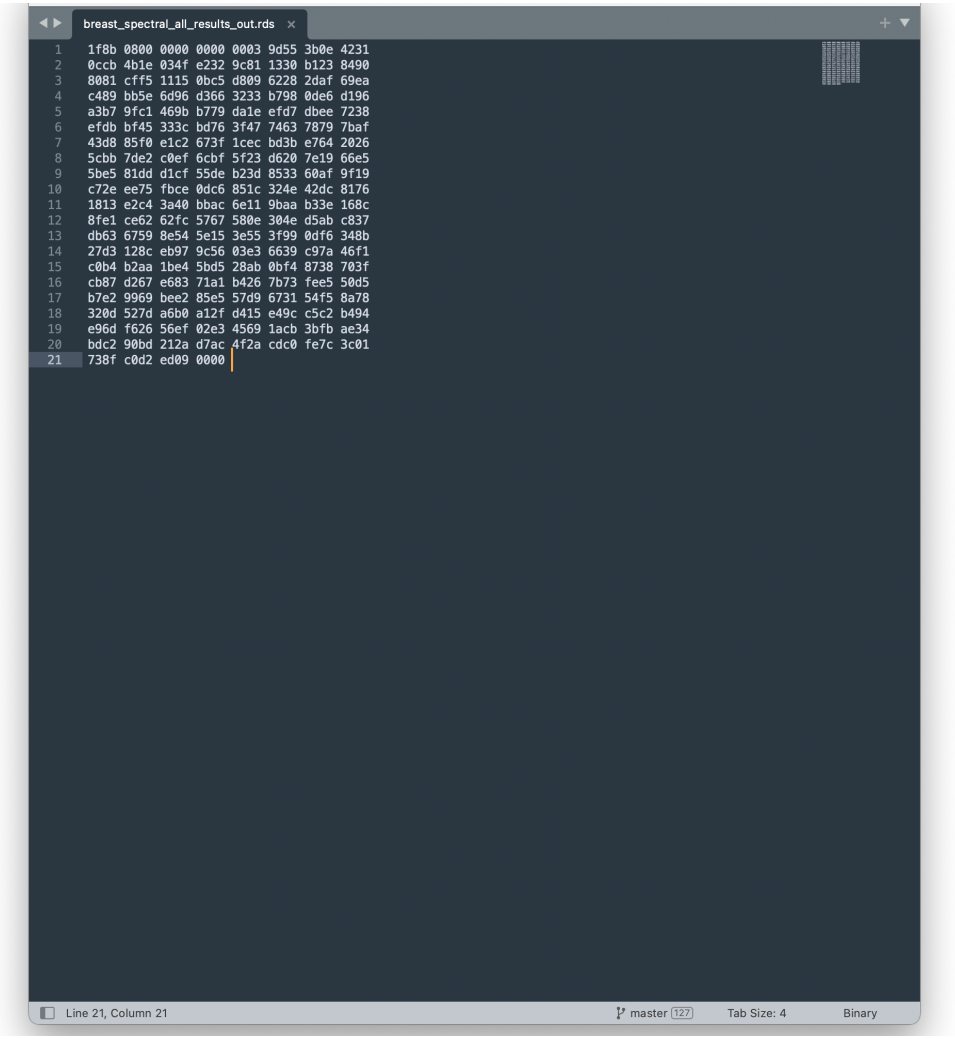
Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters.



Task 1 _b (3)

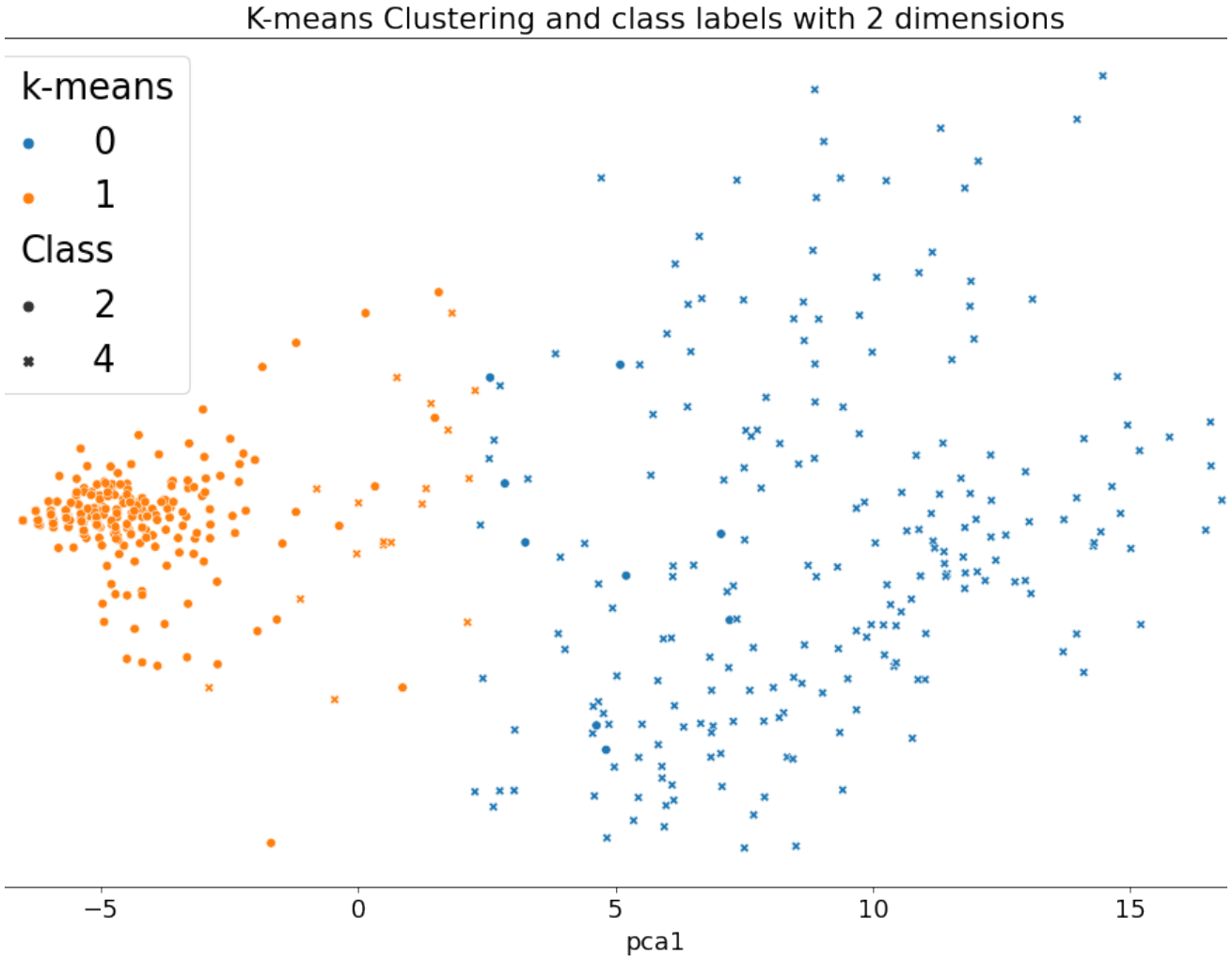
Results sample

Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters.

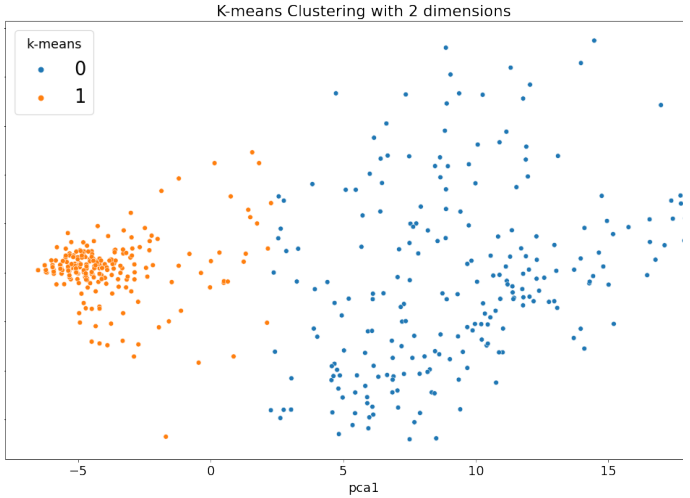


Task2

Initial classification and clustering effects

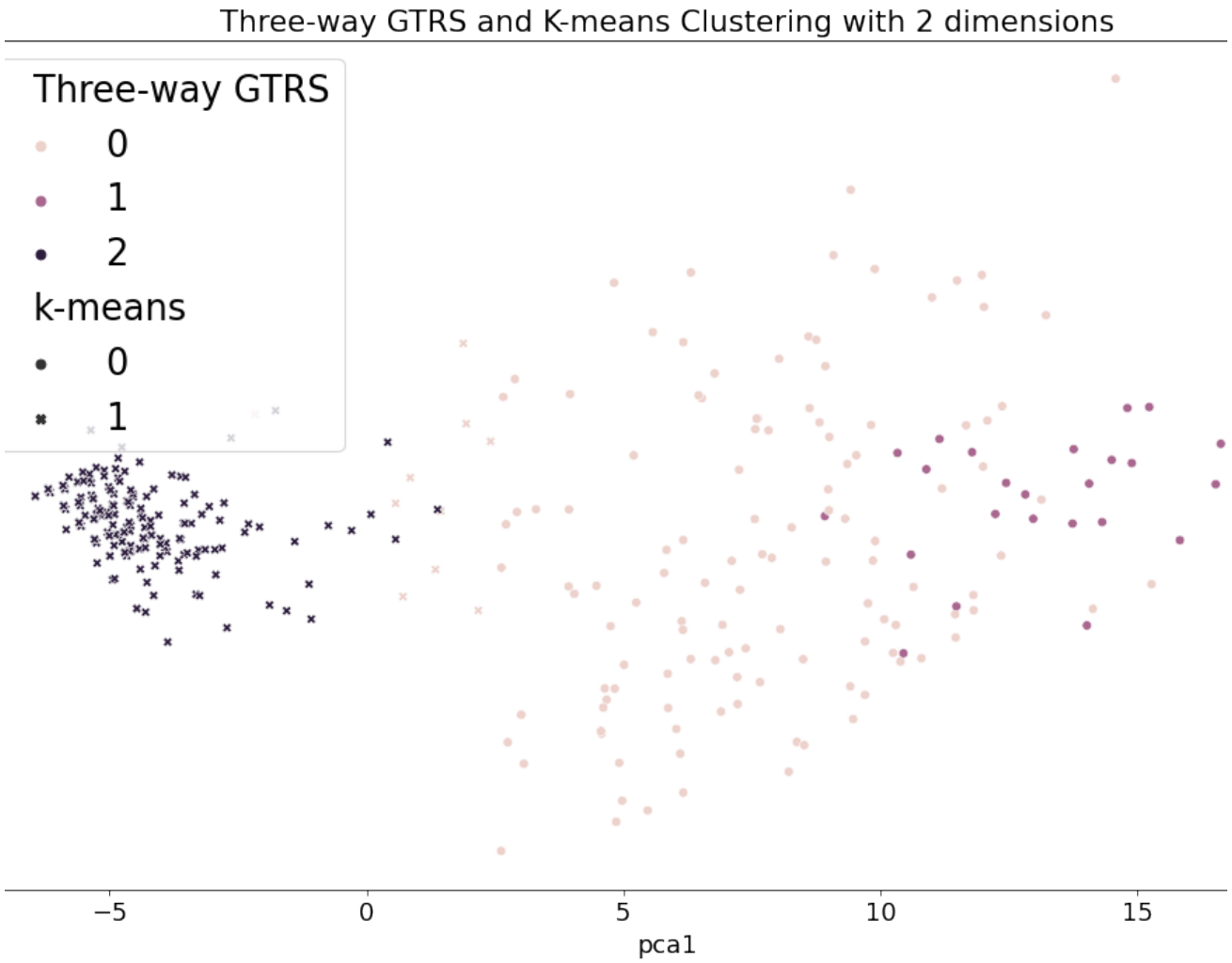

K-means Clustering and class labels with 2 dimensions
k-means
0
1
Class
2
4
−5
0
5
10
15
pca1

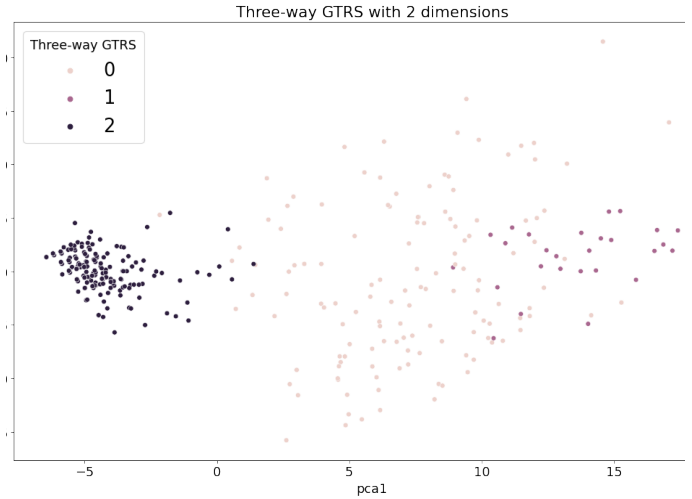

Class labels with 2 dimensions
Class
2
4
−5
0
5
10
15
pca1

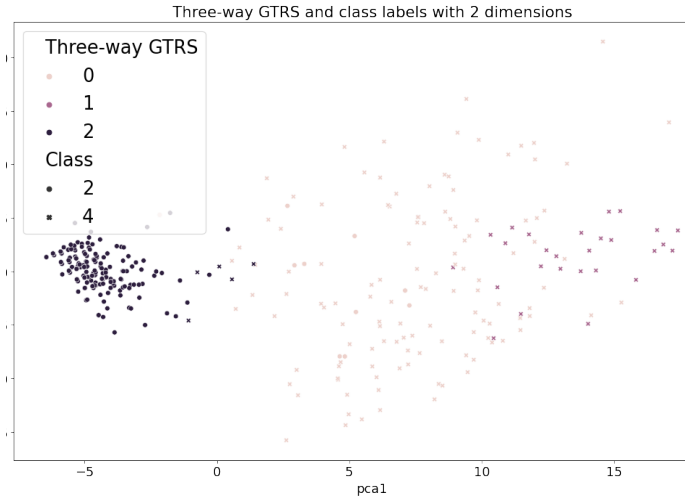

K-means Clustering with 2 dimensions
k-means
0
1
−5
0
5
10
15
pca1

Task2

Classification and clustering effects after implementing the GTRS three-way clustering algorithms


Three-way GTRS and K-means Clustering with 2 dimensions
Three-way GTRS
0
1
2
k-means
0
1
−5
0
5
10
15
pca1


Three-way GTRS with 2 dimensions
Three-way GTRS
0
1
2
−5
0
5
10
15
pca1


Three-way GTRS and class labels with 2 dimensions
Three-way GTRS
0
1
2
Class
2
4
−5
0
5
10
15
pca1

Task 2

Discussion of the limitations of the algorithm

- The GTRS three-way clustering algorithm can effectively distinguish the data distributed in the two extreme parts (inside and outside), but once the algorithm is overfitting or underfitting in the process of use, it may cause the operation time to be too long or face the large amount of data that deals with partial parts cannot be differentiated.
- To enable GTRS three-way clustering algorithm to effectively cluster data and avoid overfitting or underfitting, it is necessary to fine-tune the parameters of the algorithm (such as α-, α--, β+, β++).