Principles of Digital Biology Individual Report

11326396

1 Introduction

Clustering can help identify different subtypes of diseases by grouping patients based on similar symptoms, genetic information, or response to treatment. Visualization can help explore complex health data, detect outliers, understand distributions, identify trends or patterns, and convey findings and insights more effectively to stakeholders, including medical professionals, patients, and policymakers.

This report aims to explore the use of PCA and clustering on the dataset (recording the logged fold change in cytokine concentrations after stimulation of PBMCs by rhinovirus), and compare the results of GMM in the paper [CBL+18].

2 Methods

Steps:

1) Read the dataset and carry on Exploratory Data Analysis (EDA) based on the dataset (using the ProfileReport package).

2) As features have quite different scales, re-scaling the features using zero mean and unit variance allows them to contribute equally to the analysis and ensures that the scale of the measurements doesn't affect the algorithm's ability to learn the patterns.

3) Use PCA to reduce features' dimension, make a linear data projection into a lower dimension to observe the data distribution (2D or 3D), choose the projections that retain as much variance in the data as possible, and visualize the features driving the observed structure to explore feature importance (PCA helps to understand the clustering effect in the following parts).

4) Use the K-Means algorithm to cluster the children in the dataset, change the value of k to compare the different clustering results, and tell the difference between them and GMM results in the paper.

5) Use hierarchical clustering to cluster the cytokines, and observe the results of different linkage methods.

3 Results

Task 1:

1) Use a PCA biplot to visualize the GMM clusters in 2D and to display the five cytokines that contribute most to the first 2 principal components in **Figure 1**. The five cytokines that contribute the most are MCP4, IP-10, IL-6, IL-8, and TNF.

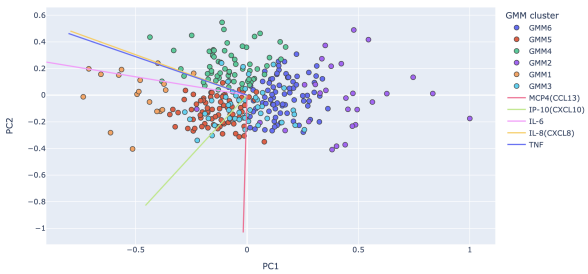


Figure 1: PCA biplot to visualize the GMM clusters in 2D

2) From **Figure 2**, only about 33.64% variance in the data is explained by the first two principal components, so the visualization is not so accurate, as the 2D plot can only explain 1/3 variance.

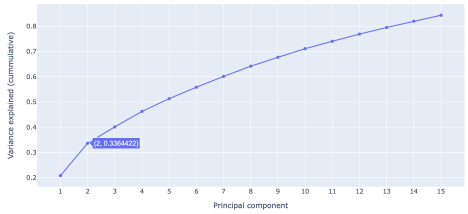


Figure 2: Variance explained (cumulative) by PCA components

Task 2:

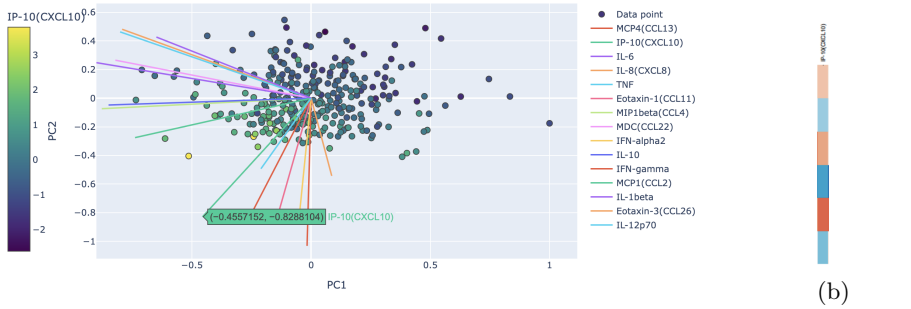
For each cytokine, use the PCA biplot to display the effect of 5 cytokines on GMM clusters, and compare them with the heatmap.

1) For MCP4, the relative intensity of MCP4 expression is GMM5, consistent with GMM5 in the heatmap in **Figure 3**, refer to **Figure 1**.

(a) biplot (b) heatmap

Figure 3: Comparison of MCP4

2) For IP-10, the relative intensity of IP-10 expression is GMM3 and GMM5(the highest), consistent with GMM5 in the heatmap in **Figure 4**, refer to **Figure 1**, .



(a) biplot (b) heatmap

Figure 4: Comparison of IP-10

3) For IL-6, the relative intensity of IL-6 expression is GMM1(the highest) and GMM 4, consistent with GMM1 and GMM4 in the heatmap in **Figure 5**, refer to **Figure 1**.

(a) biplot (b) heatmap

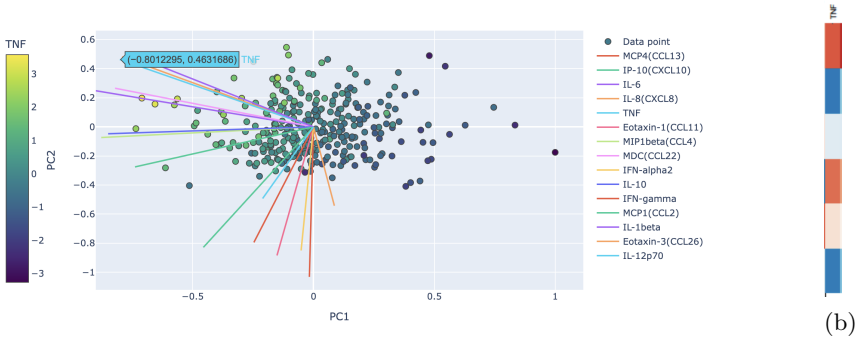
Figure 5: Comparison of IL-6

4) For IL-8, the relative intensity of IP-10 expression is GMM1 and GMM 4, consistent with GMM1(the highest) and GMM4 in the heatmap in **Figure 6**, refer to **Figure 1**.

(a) biplot (b) heatmap

Figure 6: Comparison of IL-8

5) For TNF, the relative intensity of IP-10 expression is GMM1 and GMM 4, consistent with GMM1(the highest) and GMM4 in the heatmap in **Figure 7**, refer to **Figure 1**.



(a) biplot (b) heatmap

Figure 7: Comparison of TNF

The results are consistent with the patterns observed in the heatmap (some have slight differences, like which is the highest relative intensity).

Task 3:

Use the k-means algorithm to cluster the children into 6 clusters based on their cytokine profiles and display them in **Figure 8**, which looks not great.

Figure 8: K-means results shown by PC1 and PC2

Use the adjusted rand score to measure the similarity of the two assignments, which is 0.4787. From **Figure 9**, K-Means cluster "1" agrees well with GMM4, and cluster "4" agrees well with GMM1.

Labels Clusters	GMM1	GMM2	GMM3	GMM4	GMM5	GMM6
0	0	13	0	0	0	0
1	1	0	5	59	1	3
2	0	10	2	0	1	41
3	0	2	23	1	60	1
4	17	0	0	3	2	0
5	0	3	16	0	10	33

Figure 9: Crosstable of GMM clusters and K-means clusters

Task 4:
Figure 10 shows the elbow result, it returns 5 as the optimal number of k-means clusters.

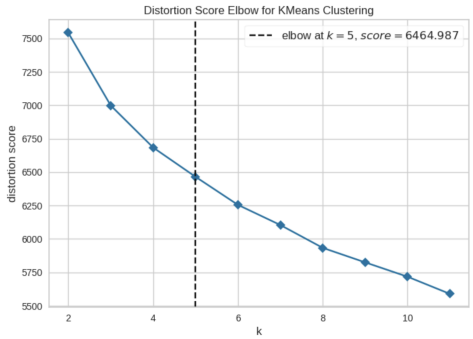


Figure 10: Elbow plot for number of k-means clusters

Figure 11 shows the distribution of 5 K-Means clusters on PC1 and PC2, which is much better than 6 K-means clusters in **Figure 8**, as the overlaps of these clusters are smaller.

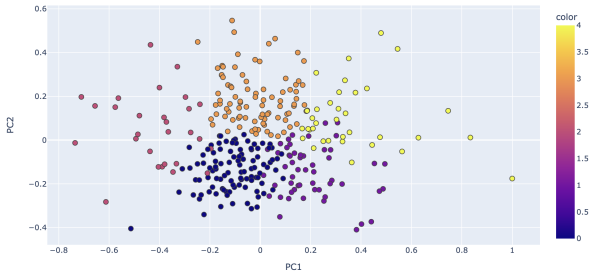


Figure 11: 5 K-means clusters by PC1 and PC2

Finally, I get the crosstable of GMM clusters and 5 K-means clusters shown in **Figure 12**, compare it with the crosstable in **Figure 9**, we can see that GMM1 cluster is most stable.

Labels	GMM1	GMM2	GMM3	GMM4	GMM5	GMM6
Clusters						
0	1	1	33	0	63	4
1	0	12	5	0	1	36
2	17	0	0	5	6	0
3	0	0	7	58	4	18
4	0	15	1	0	0	20

Figure 12: Crosstable of GMM clusters and 5 K-means clusters

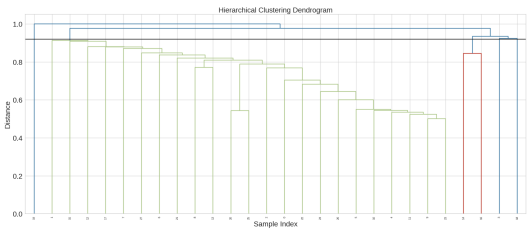
Task 5:

Use hierarchical clustering to cluster the cytokines based on their levels across children, and compare them with the cytokine categories given in the paper, the results are shown in **Table 1**. We can see the ARIs are low across the 4 methods.

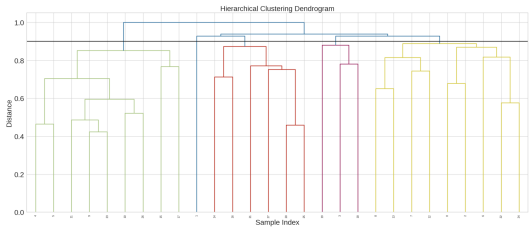
linkage methods	distance threshold	ARI(Adjusted Rand Score)
ward	0.61	0.3554
single	0.92	-0.1160
complete	0.9	0.1056
average	0.95	0.1290

Table 1: Comparison of 4 linkage methods and cytokine categories in the paper

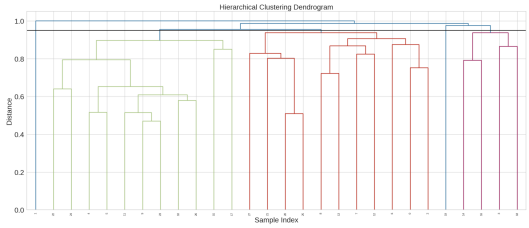
From **Figure 13**, the thresholds are all very high(over 0.09), and most clusters are joined by short horizontal lines that are highly similar, so it's difficult to separate clusters with high dissimilarity.



(a) single dendrogram



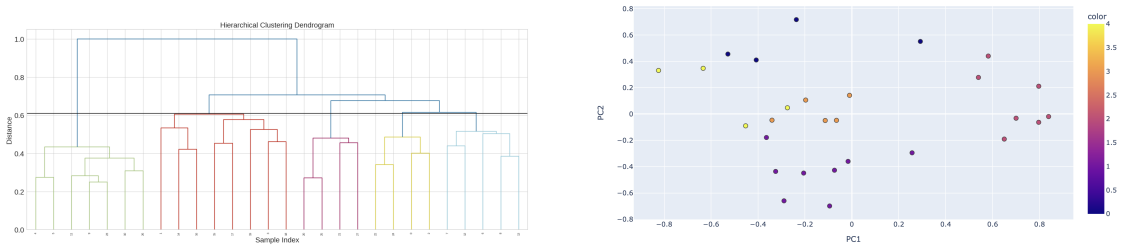
(b) complete dendrogram



(c) average dendrogram

Figure 13: Comparison of 3 linkages' dendrograms

So I choose the method "ward" with the lowest threshold to do further analysis in **Figure 14(a)**, some clusters are joined by longer horizontal lines that are not highly similar, and we can see better clustering performance in **Figure 14(b)**.



(a) ward dendrogram (b) ward linkage on PC1 and PC2

Figure 14: Hierarchical clustering with "ward" linkage

However, compared with the cytokine categories given in the figure, it doesn't agree well in **Figure 15, Figure 16**.

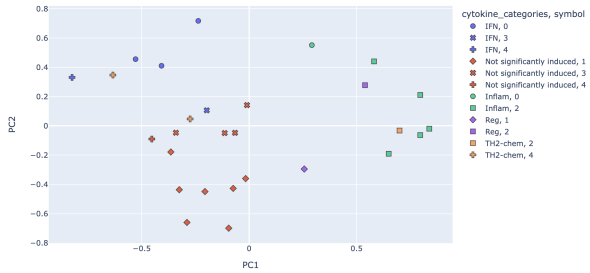


Figure 15: Crosstable of GMM clusters and hierarchical clusters with "ward" linkage

Labels Clusters	IFN	Inflam	Not significantly induced	Reg	TH2-chem
0	3	1	0	0	0
1	0	0	7	1	0
2	0	5	0	1	1
3	1	0	4	0	0
4	1	0	1	0	2

Figure 16: Crosstable of GMM clusters and hierarchical clusters with "ward" linkage

4 Discussion

From the section Result, we can see the results of K-means clustering and hierarchical clustering are not good, the reason might be the data clusters are irregularly distributed, have complex cluster shapes and different cluster sizes (in **Figure 1**), so GMM clustering is more suitable for this dataset, which the other 2 methods are not good.

References

[CBL+18] Adnan Custovic, Danielle Belgrave, Lijing Lin, Eteri Bakhsoliani, Aurica G Telcian, Roberto Solari, Clare S Murray, Ross P Walton, John Curtin, Michael R Edwards, et al. Cytokine responses to rhinovirus and development of asthma, allergic sensitization, and respiratory infections during childhood. *American journal of respiratory and critical care medicine*, 197(10):1265–1274, 2018.