Multi-view Clustering For COPDGene 10,000 Dataset

Yale Chang

1 Dataset Descritption

Dataset File: dataset_complete_knn_windsor_11-18-13.csv
Feature Annotation File: features_included_info_mhc20131104.txt
Details: There are altogether 8760 samples and 211 features in the dataset file. According to the annotation information provided in the feature annotation file, 63 continuous features(without 'Times' variables) are selected for subsequent analysis. Samples in dataset file can be divided into training set and test set according to the value of variable 'RandomGroupCode'. There are 4413 training samples. In the end, we use a dataset consisting of 4413 training samples and 63 annotated features .

2 Features Similarity matrix

2.1 HSIC

HSIC(Hilbert-Schmidt Independence Criterion) can be used to compute non-linear dependencies between features.Note that correlation coefficient can only capture linear dependencies between random variables and mutual information requires estimating the joint distribution of the random variables. While HSIC can measure dependence between random variables without explicitly estimating joint distributions. We can empirically estimate the HSIC by:

$$\mathrm{HSIC}(Z, F, G) = (n-1)^{-2}\mathrm{Tr}(K_1 H K_2 H) \tag{1}$$

where $Z = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ represents n observations of random variable X and Y, F and G are kernel spaces on X and Y respectively. $K_1, K_2 \in \mathrm{R}^{n\times n}$ are Gram matrices, $(K_1)_{ij} = k_1(x_i, x_j), (K_2)_{i,j} = k_2(y_i, y_j), (H)_{ij} = \delta_{ij} - n^{-1}$.

Since we're working with continuous features, Gaussian kernel can be used in computing Gram matrices:

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2_{HSIC}}) \tag{2}$$

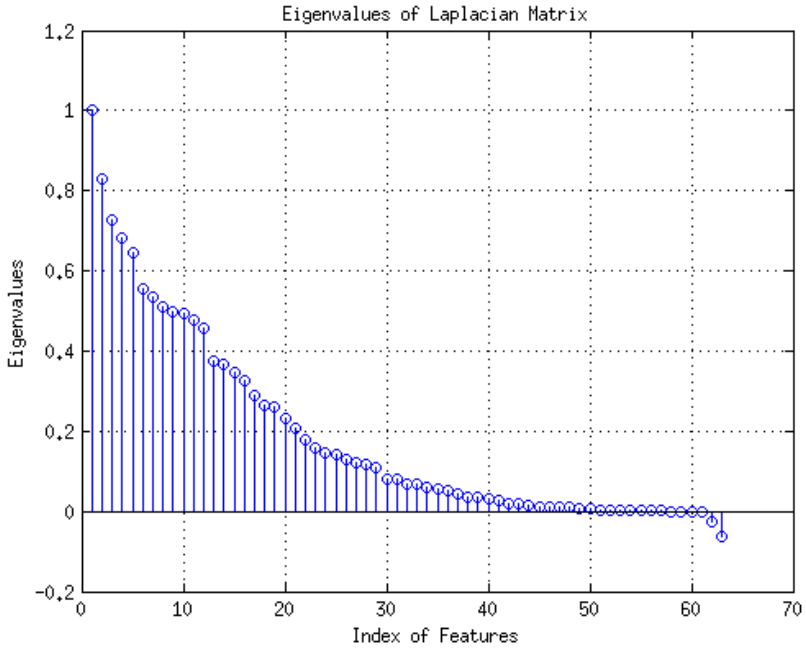

Figure 1: Eigenvalues of Laplacian Matrix

In the experiment, since the number of features is 63, we can get a 63×63 similarity matrix S_f, where $S_f(i, j)$ represents the dependency between i-th feature and j-th feature.

However, by observing equation (1), the diagonal elements of matrix S_f are not the same. To make the similarity between two features that are exactly the same to be 1, we can introduce normalized HISC:

$$\text{NHSIC}(X, Y) = \frac{\text{HSIC}(X, Y)}{\sqrt{\text{HSIC}(X, X) \cdot \text{HSIC}(Y, Y)}} \tag{3}$$

2.2 Number of Feature Clusters

Spectral clustering can be applied to cluster features. It's a common practice to determine the number of clusters by observing the eigenvalues of Laplacian matrix in spectral clustering.

From Figure(1) and Figure(2) We can observe that the top eigenvalue gaps appears at 1-2(0.1723), 2-3(0.1021), 5-6(0.0880), 12-13(0.0828). Therefore, we will set the number of feature clusters to be **2,5,12** respectively and output the plot of corresponding similarity matrix.

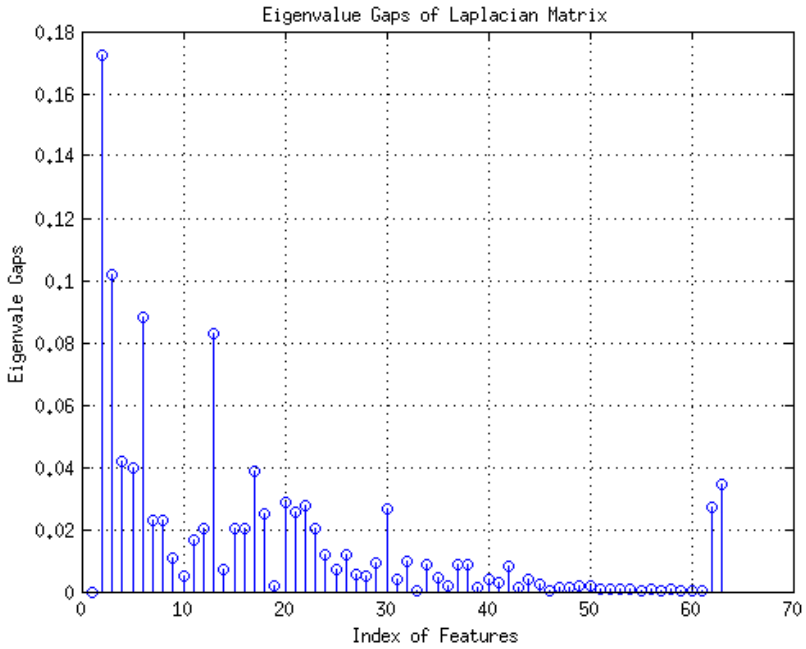

Figure 2: Eigenvalue Gaps of Laplacian Matrix

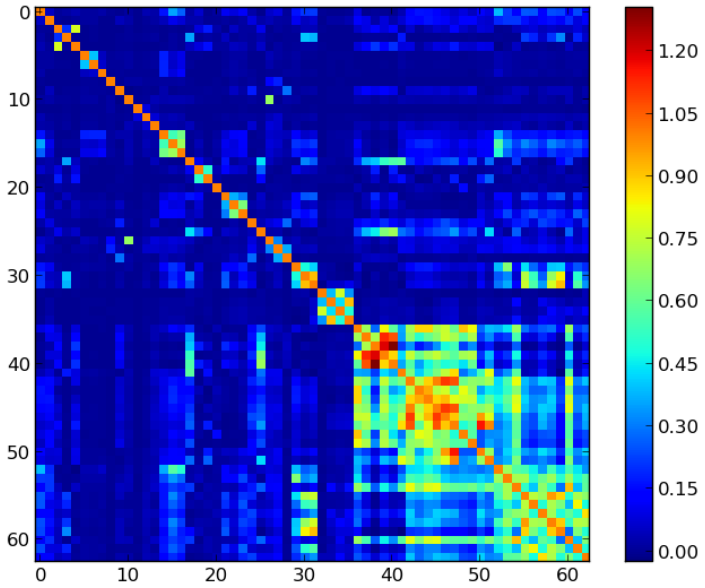

Figure 3: Similarity Matrix when Number of Feature Clusters = 2

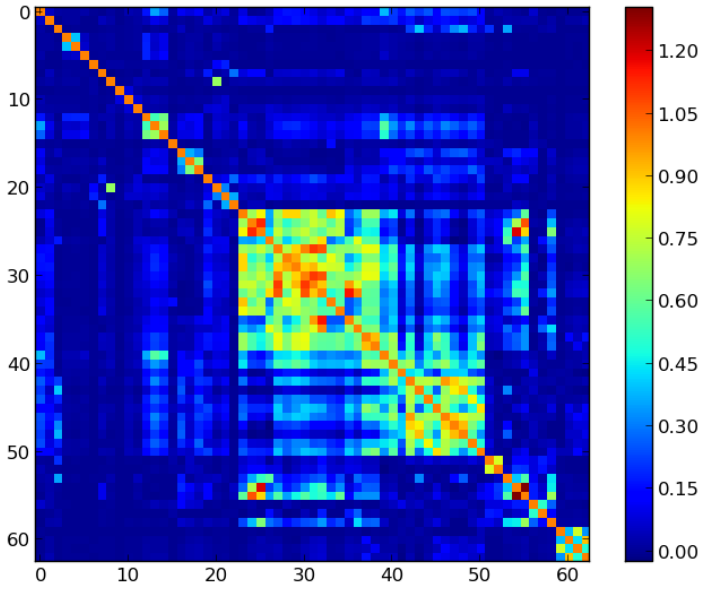

Figure 4: Similarity Matrix when Number of Feature Clusters = 5

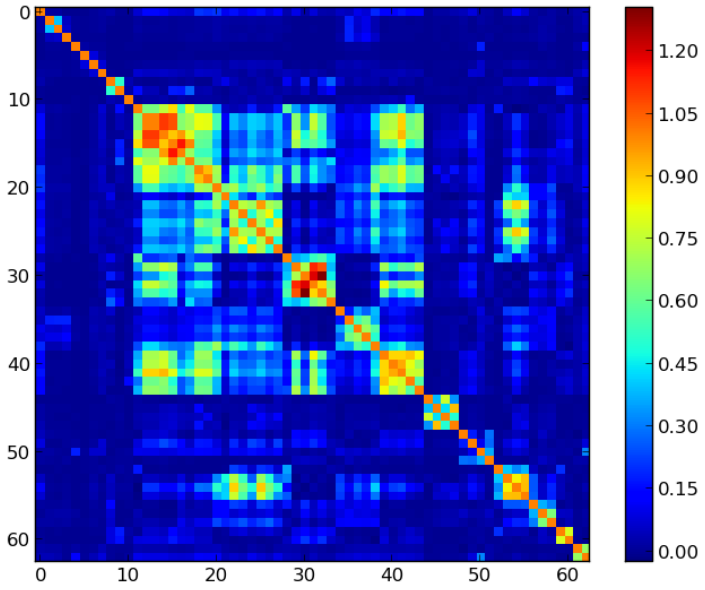

Figure 5: Similarity Matrix when Number of Feature Clusters = 12

By comparing Figure(3),Figure(4) and Figure(5), we can conclude it's more resonable to set the number of feature clusters to be 5.

2.3 Analyze Feature Clusters

When we set the number of feature clusters to be 5, the details of feature clusters are as following:

Table 1: Feature Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
distwalked	pctEmph_Slicer	BODE	Weight_KG	deltaFEV1
Resting_SaO2	Insp_Below910_Slicer	FEV1pp_utah	BMI	deltaFVC
Height_CM	Slicer_15pctIn_Total	FVCpp_utah	TLC_Slicer	BDR_pct_FEV1
CoughNumYr	FRC_Slicer	FEV1_utah	Insp_Below856_Slicer	BDR_pct_FVC
PhlegmNumYr	pctGasTrap_Slicer	FVC_utah	Slicer_IntensityStdDev_In	
NumEpisodeLastYr	Exp_Below950_Slicer	PEF_utah	Slicer_IntensityStdDev_Ex	
SmokStartAge	Exp_Below910_Slicer	PEF2575_utah	TLCpp_race_adjusted	
CigPerDaySmokNow	Exp_Below856_Slicer	pre_FEV1		
CigPerDaySmokAvg	Slicer_15pctEx_Total	pre_FEV6		
OthSmokChildYrs	Slicer_IntensityMean_Ex	pre_FVC		
OthSmokYrs	pctEmph_UpperThird_Slicer	pre_PEF		
ExpSmokAtWorkYr	pctEmph_LowerThird_Slicer	pre_PEF2575		
SGRQ_scoreSymptom	Slicer_ExpInspMeanAtten_ratio			
SGRQ_scoreActive	FRCpp_race_adjusted			
SGRQ_scoreImpact	FEV1_FVC_utah			
UpperThird_LowerThird_Slicer	pre_FEV1_FVC			
Pi10_SRWA				
Pi15_SRWA				
WallAreaPct_seg				
Age_Enroll				
ATS_PackYears				
Duration_Smoking				
YearsSinceQuit				

2.4 Clustering on Samples with Different Feature Sets

We compare the clustering results on samples with Feature Set 2, Feature Set 3 and the whole Feature Set. At the same time, we vary the number of clusters. We also visualize the clustering results under different conditions in 2D and 3D projection spaces by applying kernel PCA. We also cluster with features selected by backward search and forward search with different supervisions.The results are presented in a set of figures.