Asthma Subphenotype Classification & Clustering

DNA Microarray Data Analysis

Deepika Yeramosu | Evan Trop | Neel Mehtani

Problem Motivation & Application

Clinical Use Cases

- Observing Patient Disease Spectrums

Genomics & Precision Medicine

- Gene & Patient-Disease Pattern Learning
- Front-end Tailored Clinical Treatment

Dataset

Courtesy of: Dr. Sally Wenzel (University of Pittsburgh School of Medicine) | Dr. Wei Wu (CMU School of Computer Science)

Agilent Technologies Bioanalyzer DNA Microarray data

High Dimensional Gene Expression of BAL & BEC cells

Clustering & Classification Tasks:

- Patient Asthma Severity [0: Normal Control (NC) | 1: Not Severe Asthma (notSA) | 2: Severe Asthma (SA)]
- Hidden Gene Network Relationships Identification

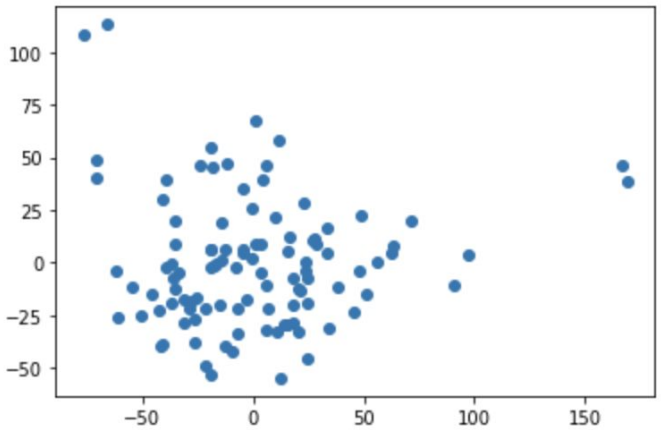
Figure 1: # of patients identified for each subphenotype in the BEC and BAL datasets

Data Preprocessing: PCA

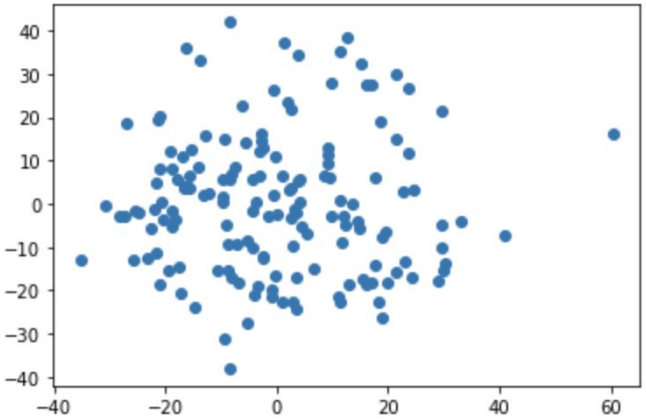
- Used PCA package from scikit-learn software in Python
- pca = PCA(.95)
  - Scikit-learn will choose the minimum number of principal components such that 95% of the variance is retained
- Original Data:
  - BEC: 155 samples, 30,889 genes
  - BAL: 104 samples, 41,000 genes
- After PCA:
  - BEC: 155 samples, 111 genes
  - BAL: 104 samples, 85 genes

PCA with 2 Components

BAL Data

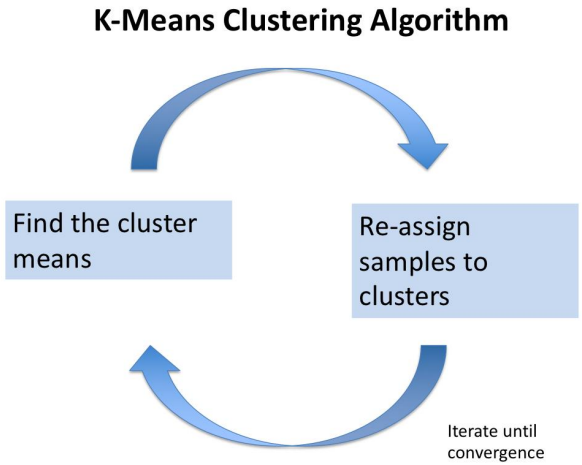


BEC Data

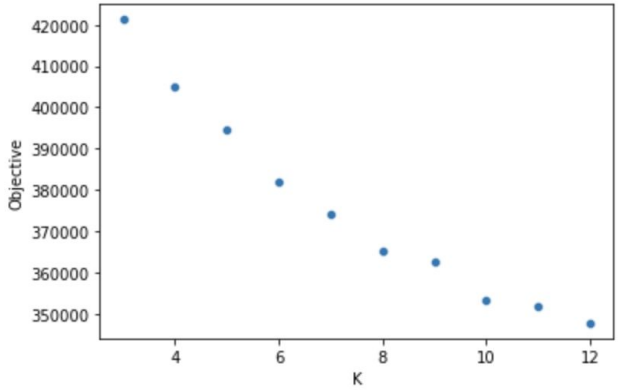


Method 1: K-Means Clustering

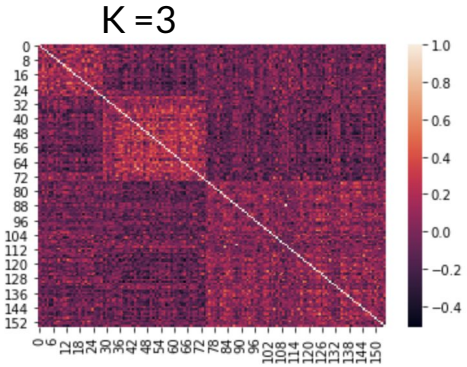
- Preliminary unsupervised learning
- Clustering of samples (patients) by features (genes)
- Each cluster represented by its mean
- Advantages:
  - Fast
  - Robust
  - Simple
  - Guaranteed convergence
- Disadvantages
  - Choosing K not always easy
  - Highly dependent on initial values
  - Difficult when number of dimensions too high
- To ameliorate disadvantages:
  - Did 10 random initializations for each K
  - PCA to reduce dimensionality

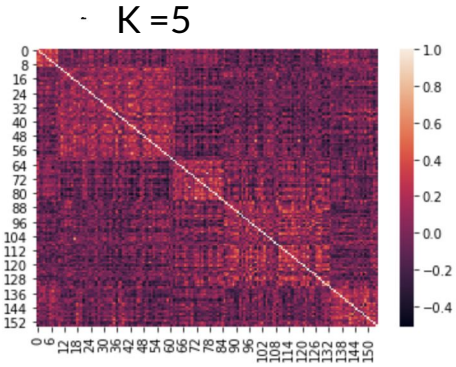


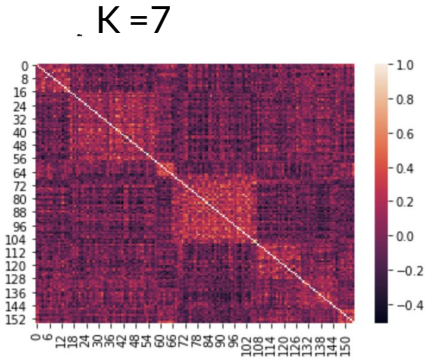
K-Means Results (BEC)

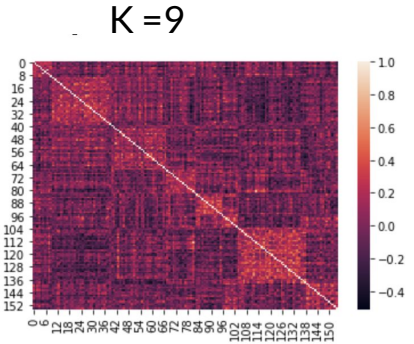


K value vs Objective









- K-Means relies on Euclidean distances
  - Only spherical shaped clusters
  - All data points have equal weight
- Sensitive to outliers

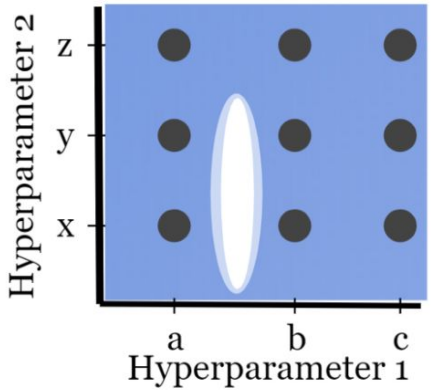
Method 2: Logistic Regression

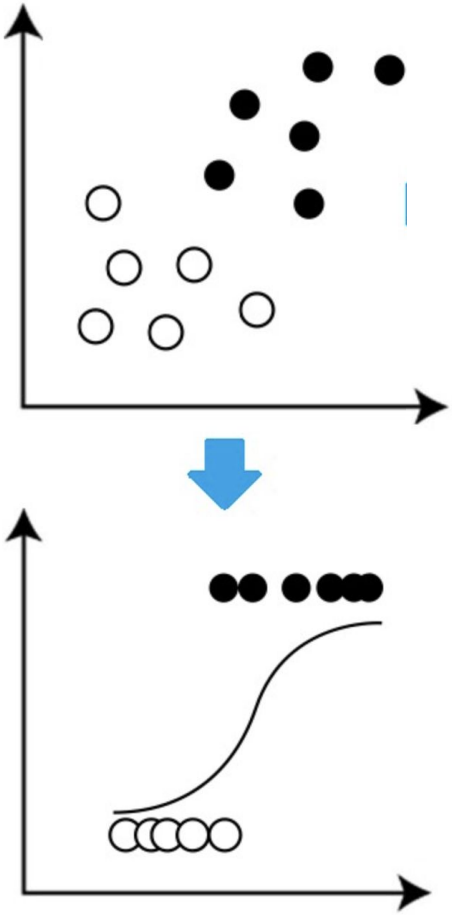
- Multiclass Classification
- X - reduced data from PCA
- Y - Normal, Mild Asthma, Severe Asthma
- Solver for multinomial loss
  - Newton-cg, lbfgs, sag, saga
- Grid Search parameter tuning
  - Penalty, C penalty strength

Grid Search

Pseudocode

```
Hyperparameter_One = [a, b, c]
Hyperparameter_Two = [x, y, z]
```

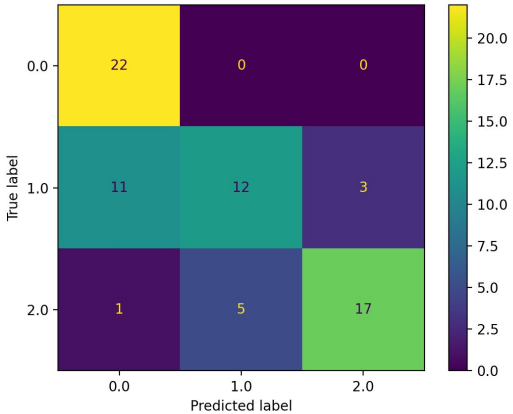




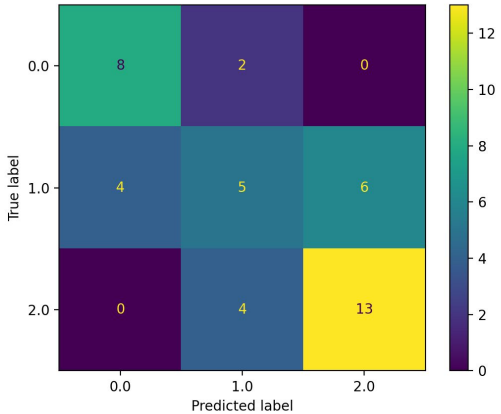
Logistic Regression Results

- 0 - Normal , 1- Mild Asthma
  2 - Severe
- Logistic regression limit to linearly separable cases
- Multicollinearity between independent variables (gene expression)

BEC Cells

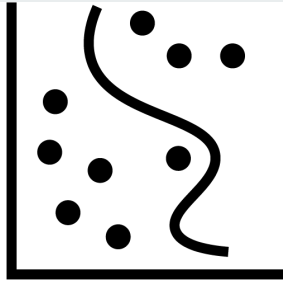


BAL Cells



Pred Accuracy: **0.72** **0.62**

Method 3: SVM



SVM Multilabel Classification for Patient Asthma Subphenotype

- 10-Fold Cross-Validation For Best Model Selection
- Resampling (over) Due to Imbalanced Data Classes

Hyperplane Segmentation of Gene Expression Data Space

- Feature Space: original (non-PCA) vs (PCA)
- Kernels Used: Linear, Polynomial, RBF
- Hyperparameters (Regularization & Polynomial Degree) Tuning

Scoring Metric

- Accuracy (with resampling)
- F1-Score (if no resampling performed)
- Would be interesting - Multiclass Confusion Matrix

SVM Results

PCA: Test Accuracy
Higher on Average

Non-PCA: Good
Bias-Variance Balance

Best Overall Kernel:
RBF

Linear Kernel (PCA) Best
Performance: 100%

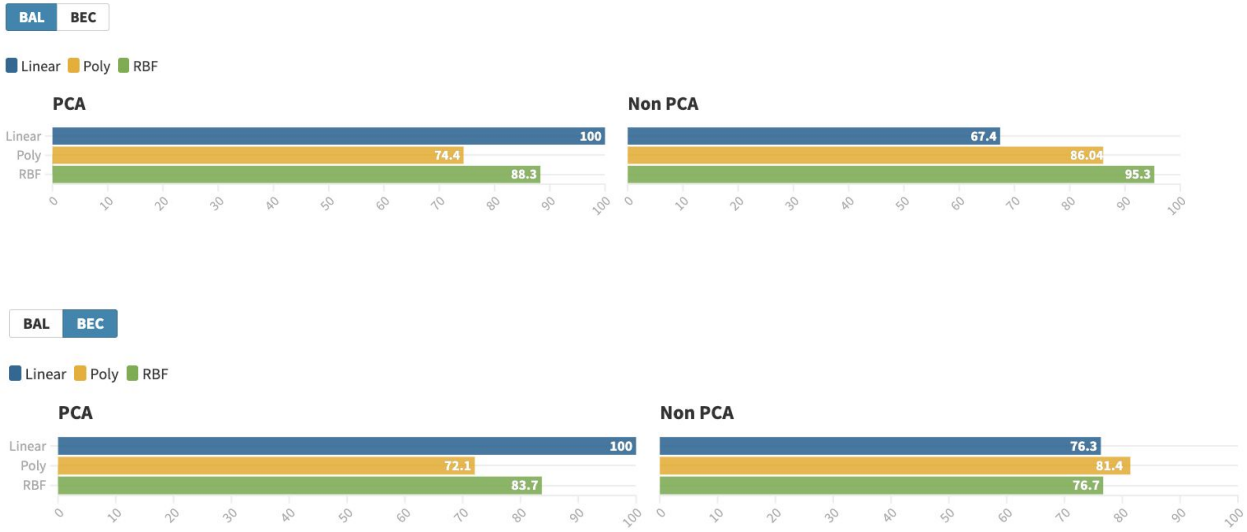


Figure : Best SVM classifier accuracy scores on PCA & non-PCA BEC & BAL test sets after hyperparameter cross-validation

Conclusion

- Working with real data vs synthetic data
- PCA and K-Means offers good preliminary analyses of data
- Explore other methods for classification
    - NB , random forest