Comparing Between Supervised and Unsupervised Approaches to Classify Gene Expression Profiles of Cancer Patients

Anna Romanov[1], Maxim Kolchinsky[1]

1 Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel

Background

- **Apoptosis** is a form of programmed cell death that occurs in multicellular organisms
- Negative regulation of apoptosis inhibits apoptosis signaling pathways, helping tumors to evade cell death.
- Identification of targets for apoptosis induction is important to provide novel therapeutic approaches in breast cancer.
- According to researchers [2], PKCδ supports surviving of breast cancer cells:

PKCδ ⊣ ERK1/2 pathway → Apoptosis

- It is therefore assumed that down-regulation of Protein Kinase Cδ can be used as treatment.

Goals

- Compare between different methods of supervised and unsupervised approaches of classifying gene expression.
- Be able to classify whether a gene is related to apoptosis based on its expression levels in different cells, with and without treatment. We expect genes associated with apoptosis show difference in expression levels between treatment and control samples, due to the role of PKCδ in suppressing apoptosis.

Gene Expression Data → Classifier → Labels (True, False)

Material and Methods

- **GEO** – public genomics data repository
- **AmiGo** – Web-based set of tools for searching and browsing the gene ontology. database. We used it to obtain a list of GO terms associated with apoptosis.
- **Scikit-learn** – Python library for machine learning.

Building the dataset

- Dataset: 12 samples of breast cancer cell lines (BT-549, MDA-mB-468) before and after down regulation of PKCδ. For each sample, gene expression levels and associated GO terms are presented.
- Generating labels:

Unite probes by genes → Apoptosis terms list from amiGO → Set label for each gene: True if associated with apoptosis term / False otherwise

Data preparation and preprocessing

- Dividing the data into training and test sets.
- Run the following process:

Normalize → Variance Filter → Clustering / Oversampling → Classifying

Tuning the parameters of filtering and oversampling:

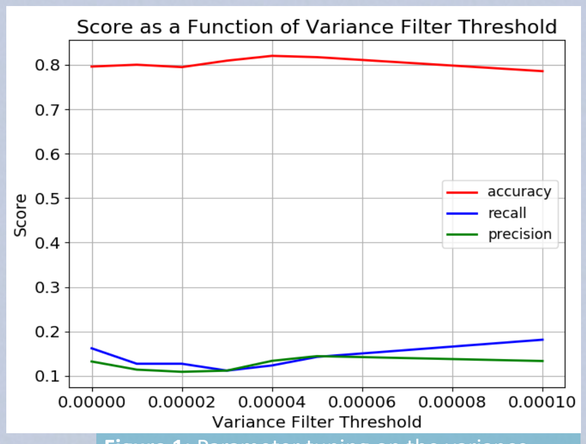

Figure 1: Parameter tuning on the variance threshold used for filtering genes out. We chose to filter out genes with variance less than 0.00005 between samples.

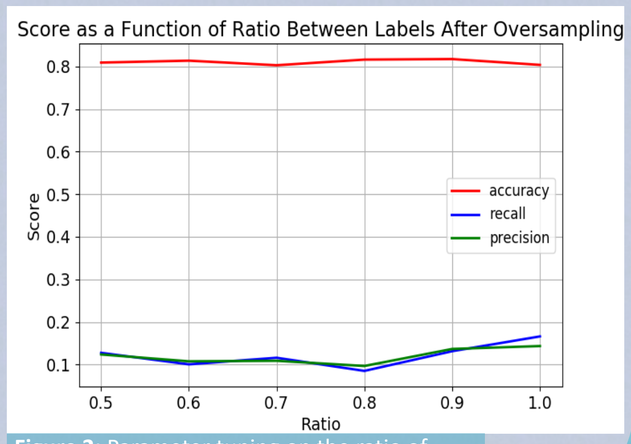

Figure 2: Parameter tuning on the ratio of 'TRUE' to 'FALSE' labels obtained after oversampling. We chose to use ratio of 1.

Classifying

Decision tree

We tuned max_depth parameter of decision tree, using 5 folds on the training set. We chose maximal depth of 70.

SVM (Support-vector machine)

Similarly, we tuned max_iter parameter of SVM classifier. We chose to allow a maximal iterations number of 20.

KNN (K nearest neighbors)

We chose to use 3 neighbors for KNN.

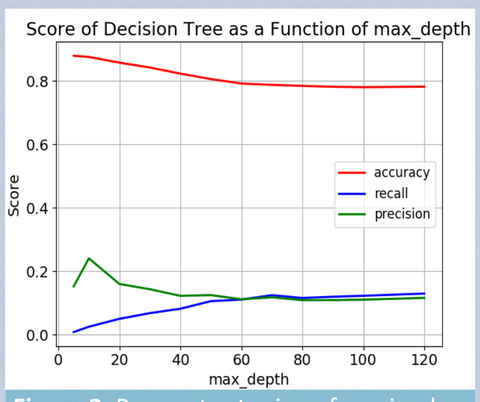

Figure 3: Parameter tuning of maximal depth allowed in decision tree classifier.

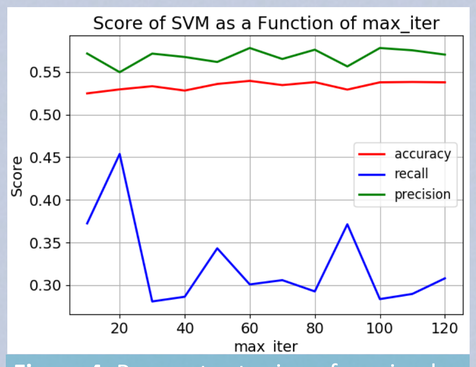

Figure 4: Parameter tuning of maximal number of iterations allowed in SVM.

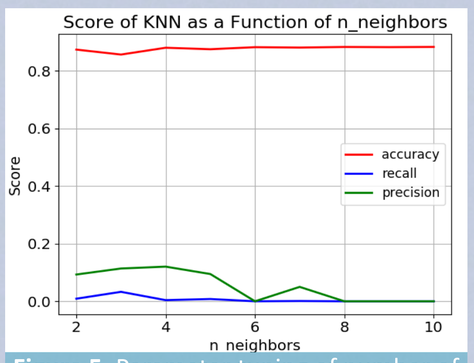

Figure 5: Parameter tuning of number of neighbors to consider in KNN classifier.

Clustering

KMeans and Hierarchical Clustering

We applied the two algorithms with Different number of clusters and compared silhouette score (based on distance within and between clusters).
We chose KMeans with 2 clusters.

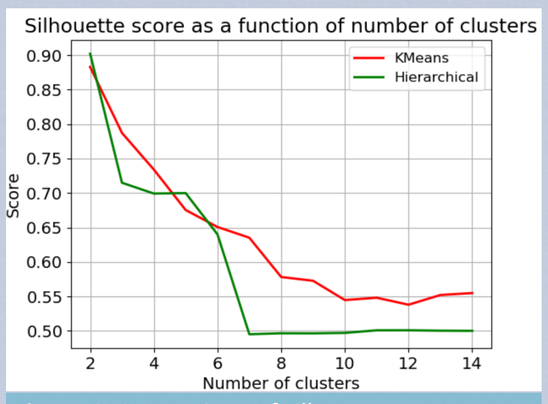

Figure 6: Comparison of silhouette score obtained for KMeans and Hierarchical clustering, as a function of the number of clusters used.

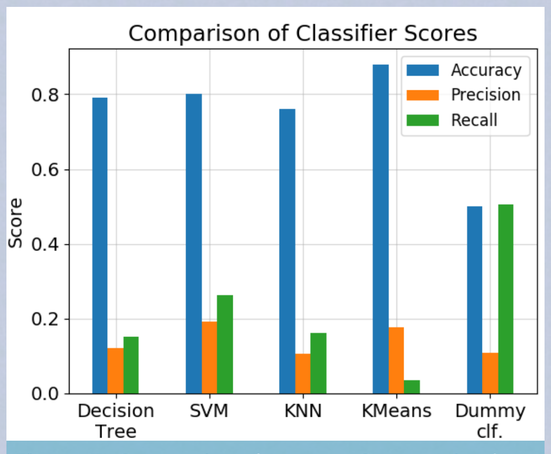

Figure 7: Comparison between 3 supervised methods, one unsupervised method and a baseline dummy classifier, considering 3 scoring methods.

Methods Comparison

We used semi-clustering for evaluation of KMeans performance (predict cluster for each sample in test, then treat clusters as labels to obtain score).
All methods show similar result in terms of accuracy, significantly better than our baseline classifier. In terms of precision and recall score we see no improvement compared to the baseline classifier.

Conclusions

- Both supervised and unsupervised methods did not show good performance.
- There was a relatively high accuracy score due to the small fraction of 'TRUE' labels in the dataset. For supervised, we managed to improve precision and recall metrics using resampling, but still most 'TRUE' genes are wrongly classified.
- For unsupervised methods, we chose the number of clusters that maximize silhouette score. However, the resulting clusters could not provide good separation of true and false labels.
- A possible explanation could be that down-regulation of PKCδ has an affect on other biological functions and therefore non-apoptosis genes also show changes in expression levels.
- Another point to note is that maybe some of the apoptosis genes are not affected by down-regulation of PKCδ and therefore their expression levels did not change, despite being labeled as 'TRUE'.
- We assume that for better results, perhaps a different labeling approach should be applied (for example, multiclass rather than binary).

References

1. **Achari C, Winslow S, Larsson C**. Down Regulation of CLDND1 Induces Apoptosis in Breast Cancer Cells. *PLoS One* 2015;10(6):e0130300.
2. **Lonne GK, Masoumi KC, Lennartsson J, Larsson C**. Protein kinase Cdelta supports survival of MDA-MB-231 breast cancer cells by suppressing the ERK1/2 pathway. The Journal of biological chemistry. 2009;284(48):33456–65