Clustering of Tumor

Monday, March 30, 2020 5:53 PM

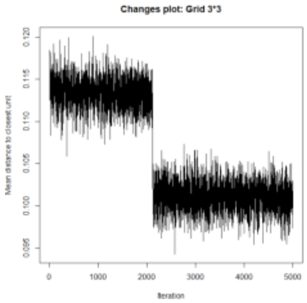
1. Clustering of Tumor Micro-array data using Self-Organizing Maps

Observations :

- This is a Human Tumor Microarray Data
- There are 6830 samples and 64 variables. The 64 variables represent results from different cancer tests, where there are 14 unique tests
- The data is for 14 subtypes of tumor cells and thus we expect 14 different clusters from our result.
- We will transpose the dataset. Now we see that we have 64 data points having 6830 predictors for each and we wish to classify the clusters into 14 groups.
- The min and max of 64 variables are varying so we can begin with scaling the dataset

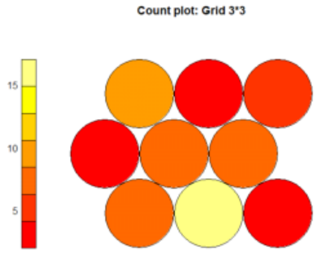
Fitting a Self-Organizing Map :

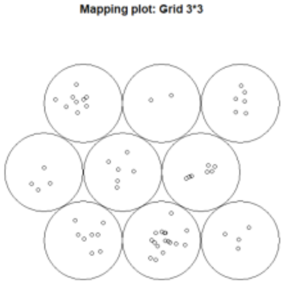
- Self-Organizing map gives us the benefit of finding the clusters when we do not know the number of clusters present.
- It also helps us in visualizing the data in 2-dimensions
- As the unique number of variables are 64, let's begin with a Grid size of 3*3 and run for 5000 iterations. It is seen to stabilize around 2500 iterations. We can conclude that the algorithm has converged



Let's see how the cancer tests have been grouped. The below plot shows the distribution of points across the prototypes in SOM.

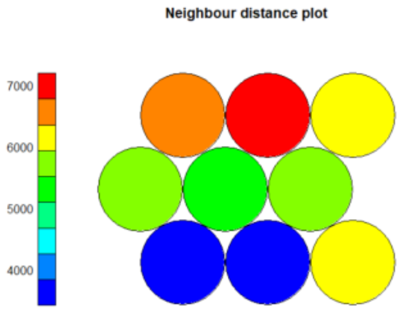
There are 64 data points and almost one-fourth of them are part of one prototype when we make a grid of 3*3





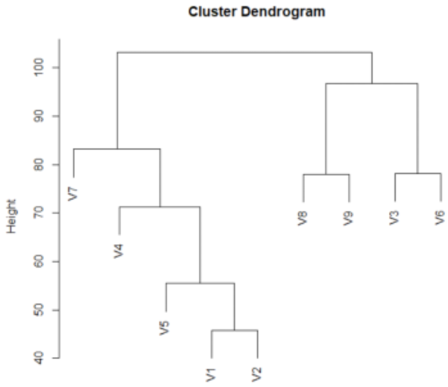
U-Matrix

- The below plot shows the sum of the distances to all immediate neighbors.
- Units near a class boundary can be expected to have higher average distances to their neighbors.
- The prototypes colored in Blue are more similar to their neighbors and the ones colored Red are dissimilar to their neighbors.
- Higher the distance, the cluster is more unique
- The U-Matrix helps us to see the quality of clusters formed
- We can see that there is one prototype that is very different (red) and few which are closely connected (in dark blue)
- Orange and Yellow ones are closer to Red whereas the Green ones can be seen to be closer to Blue.
- We can expect 2 or 3 clusters from the data. Let's explore the dendrogram to get more clarity on number of clusters



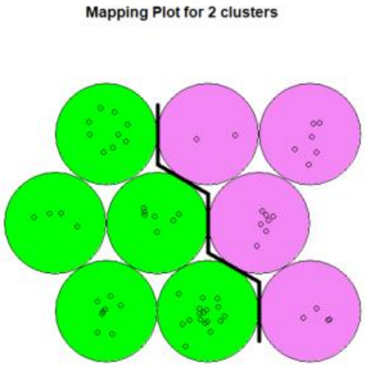
Hierarchical clustering using complete linkage and Euclidean distance measure to cluster the 9 prototypes

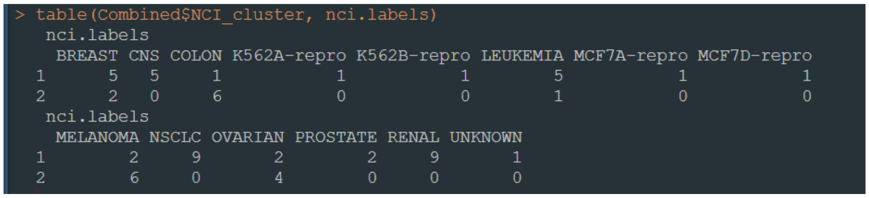
- □ The above 9 prototypes are represented as 9 datapoints and we further perform hierarchical clustering on them
- □ The results are below



- Looking at the dendrogram, division into 2 clusters looks like a natural choice.
- These results are based on Complete Linkage. The other kinds of linkage did not give good results.
- We can cut the above dendrogram such that we get 2 clusters

Self-Organizing Map Grid along with results from Hierarchical Clustering





```
> table(Combined$NCI_cluster, nci.labels)
   nci.labels
    BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro MCF7D-repro
  1      5   5     1           1           1        5           1           1
  2      2   0     6           0           0        1           0           0
   nci.labels
    MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
  1        2     9       2        2     9       1
  2        6     0       4        0     0       0
```

Inference :

- On the SOM grid, we further mark the two clusters found using hierarchical clustering with complete linkage
- We see from the final clusters obtained that one of the cluster has all NSCLC, Prostrate, Renal, MCF7D-repro, MCF7A-repro, K562B-repro, K562A-repro.
- The other cluster majorly contains most of Melanoma, Ovarian, Colon