# Self-Organizing Map for Clustering

Monday, March 30, 2020 5:53 PM

## Comparison of Clustering Algorithms : Hierarchical Clustering and Self Organizing Maps

### Data Source : USArrests data

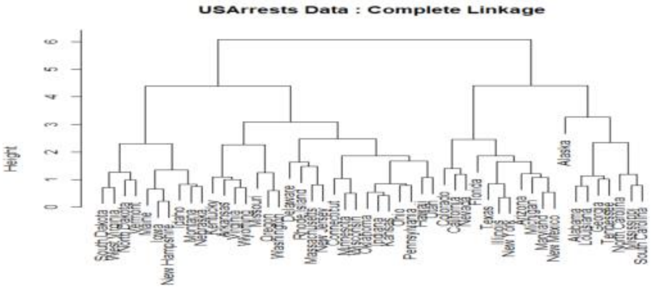
### Observations :

- There are 4 different kinds of observations - murder, Assault, Urban population and Rape for all 50 states of USA.
- This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas

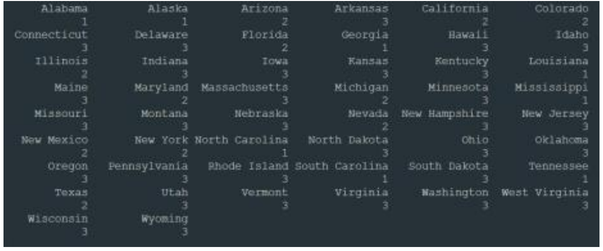
### Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

Steps :

1) The metrics are in different scales so we need to scale the data first before performing any analysis
2) Find dissimilarity matrix for the dataset.
3) Clustering the US states using complete linkage and Euclidean distance.



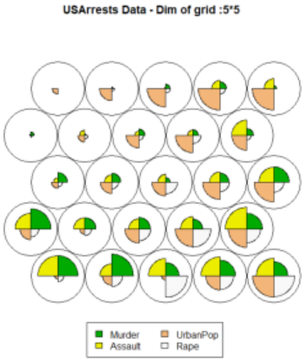
1) Cut the dendrogram at a height that results in three clusters.



```
       Alabama         Alaska        Arizona       Arkansas     California       Colorado
             1              1              2              3              2              2
   Connecticut       Delaware        Florida        Georgia         Hawaii          Idaho
             3              3              2              1              3              3
      Illinois        Indiana           Iowa         Kansas       Kentucky      Louisiana
             2              3              3              3              3              1
         Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi
             3              2              3              2              3              1
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey
             3              3              3              2              3              3
    New Mexico       New York North Carolina   North Dakota           Ohio       Oklahoma
             2              2              1              3              3              3
        Oregon   Pennsylvania   Rhode Island South Carolina   South Dakota      Tennessee
             3              3              3              1              3              1
         Texas           Utah        Vermont       Virginia     Washington  West Virginia
             2              3              3              3              3              3
     Wisconsin        Wyoming
             3              3
```

### Self-Organizing Maps

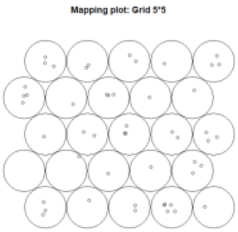
**SOM is created with Grid size 5*5 and run for 3000 iterations.**



### Inference :

- In some prototypes, we see Urban Population dominating, in some others we see the crimes dominating.
- Since there are 50 states, and we wish to cluster the states in 3 clusters (same number of clusters as hierarchical) for comp arison, let's make 25 prototypes and divide the states.
- The algorithm has run for 3000 iterations and is seen to stabilize around 2500 iterations.

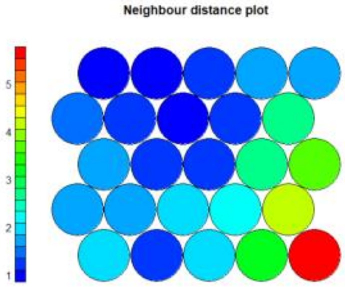
**The below plot shows the distribution of points across the prototypes in SOM.**



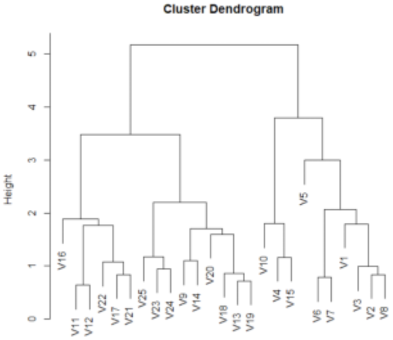
- Since there are 50 states and 25 prototypes, we have lot of protypes with 1 states and few with none.
- There are couple of protypes with 4 states, showing similarity between these states

## U-Matrix

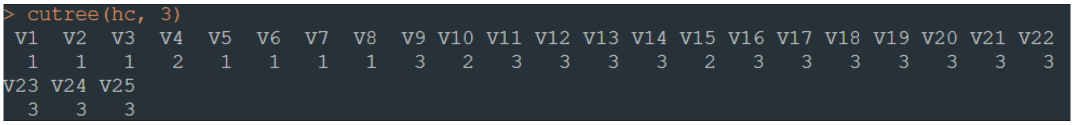
- The below plot shows the sum of the distances to all immediate neighbors.
- Units near a class boundary can be expected to have higher average distances to their neighbors.
- The prototypes colored in Blue are more similar to their neighbors and the ones colored Red are dissimilar to their neighbors .



**Hierarchical clustering using complete linkage and Euclidean distance measure to cluster the 25 prototypes into 3 clusters**

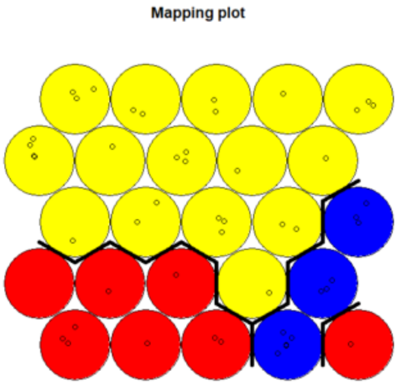


**Cutting the dendrogram to get 3 clusters.**



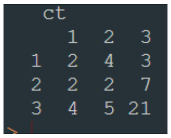
```
> cutree(hc, 3)
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 
  1   1   1   2   1   1   1   1   3   2   3   3   3   3   2   3   3   3   3   3   3   3 
V23 V24 V25 
  3   3   3 
```

**Taking the clusters back to the SOM Grid and creating the cluster division in the Grid**



## Comparing the Clustering Results :

- Using the RAND index measure to compare the clustering results from hierarchical clustering with 3 clusters and Self-Organizing Map along with Hierarchical Clustering
- RAND index measures are between 0 and 1. If the measure is 0 then the 2 clustering is complete opposite whereas when the measure is 1, the clustering is exactly the same.
- Rand Index Measure : 0.54, Adjusted Rand Index : 0.078
- The Rand index measure shows that the clustering obtained from the two methods are very different.



```
   ct
     1  2  3
  1  2  4  3
  2  2  2  7
  3  4  5 21
```

## Inference :

**When to prefer Hierarchical Clustering?**

- When the number of clusters are known to us in advance and we know that there is a hierarchical structure in data
- The grouping is done sequentially and we can choose between different kinds of linkages and distance metrics to use.

**When to prefer Self-Organizing Map?**

- Self-Organizing Maps are helpful helps us to visualize the data in 2-dimensions.
- This technique is useful when the number of clusters are unknown in the beginning.