Analysis Report for KNNC and K Mean Cluster

Analysis Report KNNC Cluster:

The Data set of engineering students have following features obtained from the web with addition of few subject fields. It has 1000 Patterns and 25 features.

gender
race/ethnicity
parental level of education
lunch
test preparation course
m1 score
m2 scorep
physics score
m3 score
m4 score
DS score
CO score
LD score
SOM score
ED score
OS-2 score

etc..

The non-numerical features like "gender", "race/ethnicity", "lunch" etc are removed. Then the KNCC Leader Algorithm implemented is run on the data points.

- Euclidean distance between cluster leader and DP is > threshold, Data points (DP) associated with different clusters.
- Number of clusters created depends on the data points euclidean distance from the leaders and threshold value. Shuffling the DPs leads to different leaders formed. Classification depends on DPs Position as well as theta . Numbers of clusters created is not equal to threshold values.
- theta/ Threshold <= { all data points euclidean distance} , then no. of clusters created = no.of data points . Here the error is ZERO.

  Average Case:

Consider theta = 100 based on the experiments done for cluster the DPs.

If theta = 100 - average euclidean distance around 100 to 140, same data points are mapped to more number of clusters. [soft clustering]

Experimented using different Theta values [35, 100, 200] and found most of the values around [100 to 140].

- Number of Clusters created is DYNAMIC , not known prior. It is determined based on the DPs and theta, euclidean distance of DPs from the data points.

  Example:

  No. of new clusters for the 1000 patterns are [71] for index 4

- **KNNC is** computation intensive and it is based on theta. If the theta is small, I have seen many clusters/leaders created and DPs euclidean distance should be calculated with all these leaders - leading to slow convergence.

  Example :

  Cluster List : [0, 1, 2, 3, 8, 9, 10, 12, 13, 14, 18, 22, 23, 24, 33, 35, 39, 44, 46, 47, 53, 55, 57, 69, 74, 76, 83, 94, 98, 104, 107, 108, 121, 124, 132, 136, 144, 148, 151, 170, 210, 232, 276, 277, 284, 307, 310, 359, 373, 401, 417, 430, 434, 465, 467, 475, 513, 527, 535, 567, 677, 752, 775, 799, 851, 886, 892, 905, 914, 925, 981]

- The Error (MSE) for different shuffles Data set is given below: It is clearly seen, **number of clusters created with the same (theta) changes due to shuffling of Data Points.**

  ---------- K NNC Table - Error(MSE) ---------->

  Index 1 | Error 4282358.60 | No of cluster: 75 | theta 100

  Index 2 | Error 3866721.24 | No of cluster: 73 | theta 100

  Index 3 | Error 3959840.80 | No of cluster: 74 | theta 100

  Index 4 | Error 4022552.80 | No of cluster: 74 | theta 100

  Index 5 | Error 4086689.25 | No of cluster: 74 | theta 100

  Index 6 | Error 4223218.69 | No of cluster: 67 | theta 100

  Index 7 | Error 4011023.12 | No of cluster: 79 | theta 100

  Index 8 | Error 3938759.62 | No of cluster: 79 | theta 100

Index 9 | Error 4199622.78 | No of cluster: 71 | theta 100

Index 10 | Error 4160122.57 | No of cluster: 69 | theta 100

==========

Analysis Report K Mean Cluster:

The Data set used is from web - engineering of the students , with addition of few engineering subject fields

gender
race/ethnicity
parental level of education
lunch
test preparation course

m1 score
m2 scorep
physics score
m3 score
m4 score
DS score
CO score
LD score
SOM score
ED score

OS-2 score

etc.. Procedure Followed:

The non-numerical features like "gender", "race/ethnicity", "lunch" etc are removed. Then the KNCC Leader Algorithm implemented above is run on the data points.

---------- K Mean Table - Error(MSE) ---------->

K-Mean(2), ======> 6828085.83

K-Mean(5), ======> 6320805.80

K-Mean(10), ======> 5849035.98

K-Mean(50), ======> 4554452.91

K-Mean(100), ======> 3890762.91

K-Mean(1000), ======> 0.00

- K-Means is relatively an efficient method.However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminate at a local optimum.
- Large k probably decreases the error but increases the risk of overfitting.
- Shuffling the DP do not change the Centroids after convergence

Example :

- **Error values are reduced as we increase the clusters (K) as shown above.**
- Error is dependent on the clusters(K value). For 1000 different DP, we have 1000 clusters (ie K =1000) then Error is ZERO.
- For K = 100, the Error is **3890762.91 and K = 50 the error is 4554452.91. So Error is decreasing when we increase the K value.**
- If K is small - error is HIGH and Error decreases as we increase the K values.
- For the given data set, it was observed that K Mean computation is faster than KNNC because first centroids are found and then classified wrt centroids whereas KNNC - all DPs should be validated against all the Leaders.
- Cluster creation - we need to specify the K value, number of cluster to be created, this is not the case with KNNC

END OF ANALYSIS REPORT
THANK YOU