Monte Carlo Health Attuned Multiple Metrics Evaluation Rubric - preliminary tests

Nonie

09/03/2020

Introduction

Clusters can be evaluated by comparing their distribution to that of a null distribution, however several parameters in that need to be tested first, for example what statistic to compare the original data set and the null distribution generation. Below we start the initial testing of those parameters on simple known cluster data sets.

Method

Null Distribution

Three methods were chosen to create null distributions from the data

1. **Shuffle the data** takes all the original data points and shuffles the order in the variables to remove correlation between the variables
2. **Max Min Uniform Distribution** is generated from a uniform distribution between the minumum and maximum values of the variable
3. **PCA Distribution** takes the eigan vectures of the data set are gained through PCA, these are then used to transform a random data set generated from a single gaussian distribution. The resulting data set is one with only one cluster yet maintains the relationships between the variables

To create a null distribution, 500 test data sets were generated

Cluster Seperation Metrics

Three Seperation metrics are used:

1. **Huberts Gamma Statistic** is a measure of how much the high distances between variables correlates with cluster membership. It uses 2 matrices the distance matrix which was the basis of clustering (D) and a matrix recording cluster membership where the value at point (i,j) is 1 if they are from different clusters and 0 if they are from the same. The statistic = the sum of D(i,j) * C(i,j) for i in 1-n and j in 2 -n / the number of point pairs. The higher the value the better cluster structure
2. **Normalised Gamma Statistic** is the normalised version of the statistic above. The statistic = (the sum of D(i,j)-mean(D) * C(i,j)- mean(C) for i in 1-n and j in 2 -n / the number of point pairs)/var(D)*var(C). This returns a value between 0 and 1 with high being more clustered
3. **Total Within Cluster Sum of Squares** is the sum of the distances from each point to its assigned cluster center, the smaller the distance the better.

Cluster Methodology

We apply k-means to each data set using a k++ initialisation with 50 resamples which then returns the optimum result

Test Data Generation

The data was generated using SciKit learn Make Classifications function from the datasets module. 4 parameters of the data are altered we used a full factorial experimental design:

1. Number of clusters - 2,4,5
2. Number of features - 10,20
3. % Noise features - 0% 10% 50%
4. Seperation (measured in size of hypercube between clusters) - 0.5,1,3

This resulted in 54 distinct data sets.

Overall Experiment Structure

1. 54 Datasets were created
2. For each data set 500 null distributions were made with each null distribution method (total 1500 null distributions
3. K-means was run on the original data set and 1500 null datasets and the three cluster seperation metrics were returned, for k = 2-6
4. The mean and standard deviation is returned for each null distribution method, for each seperation metric and for each cluster number
5. The seperation metric score and p value for the original data set is returned for each null distribution method, for each seperation metric and for each cluster number

Experiment Outcomes

Each distribution method and Cluster metric will by the accuracy of identifying the correct cluster number

##Results

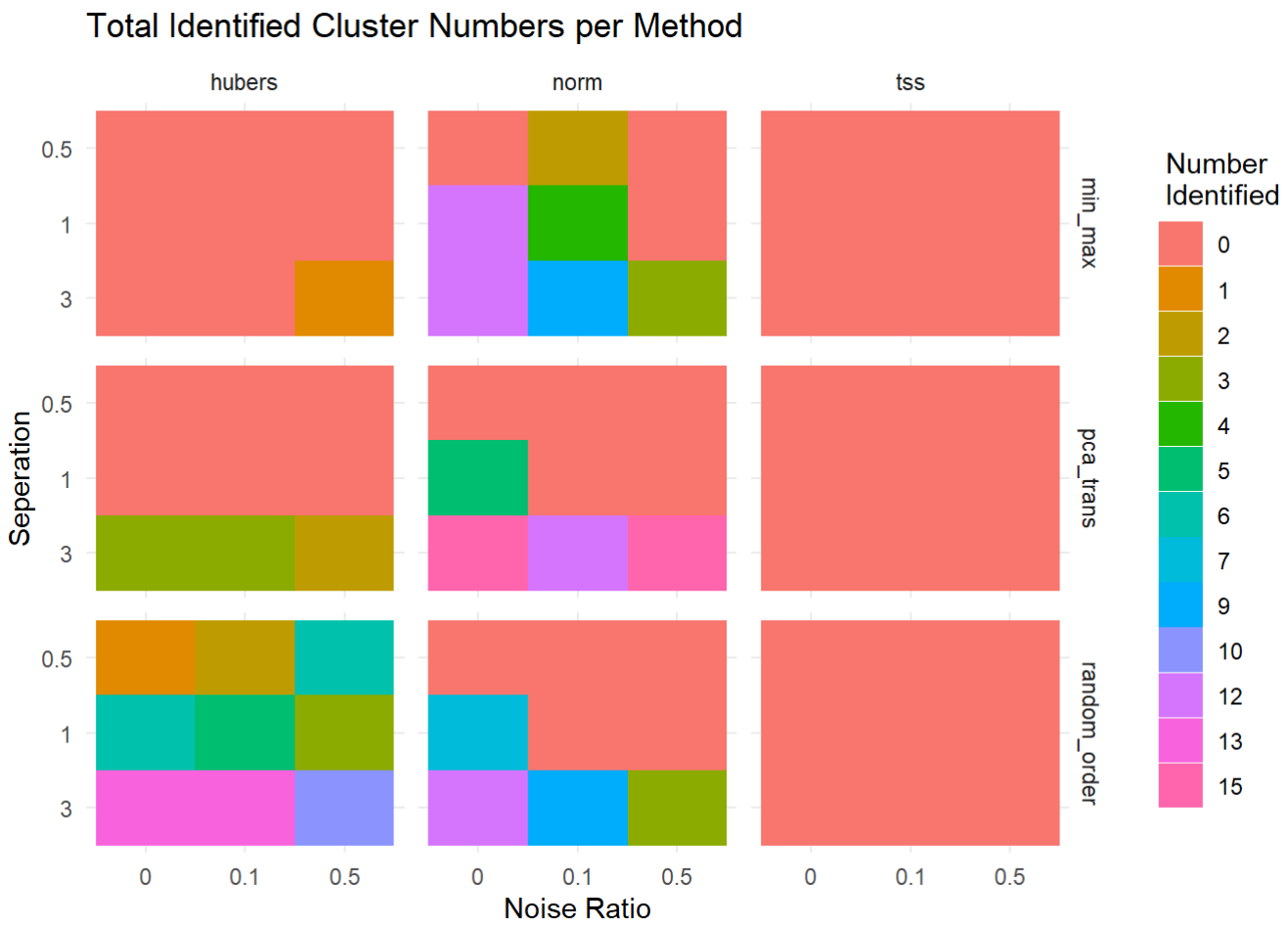
Results

Figure 1 shows the senstivity for each metric, null distribution combination. The first thing to note is that using the within sum of squares was unsuccsesfull no matter what the distribution method used. The best method used was the combination of random order generation and huberts gamma statistic with a sensitivity of .5 which is still pretty bad. Overall out of the data genration methods random order performed the best, followed by pca then lastly min max.

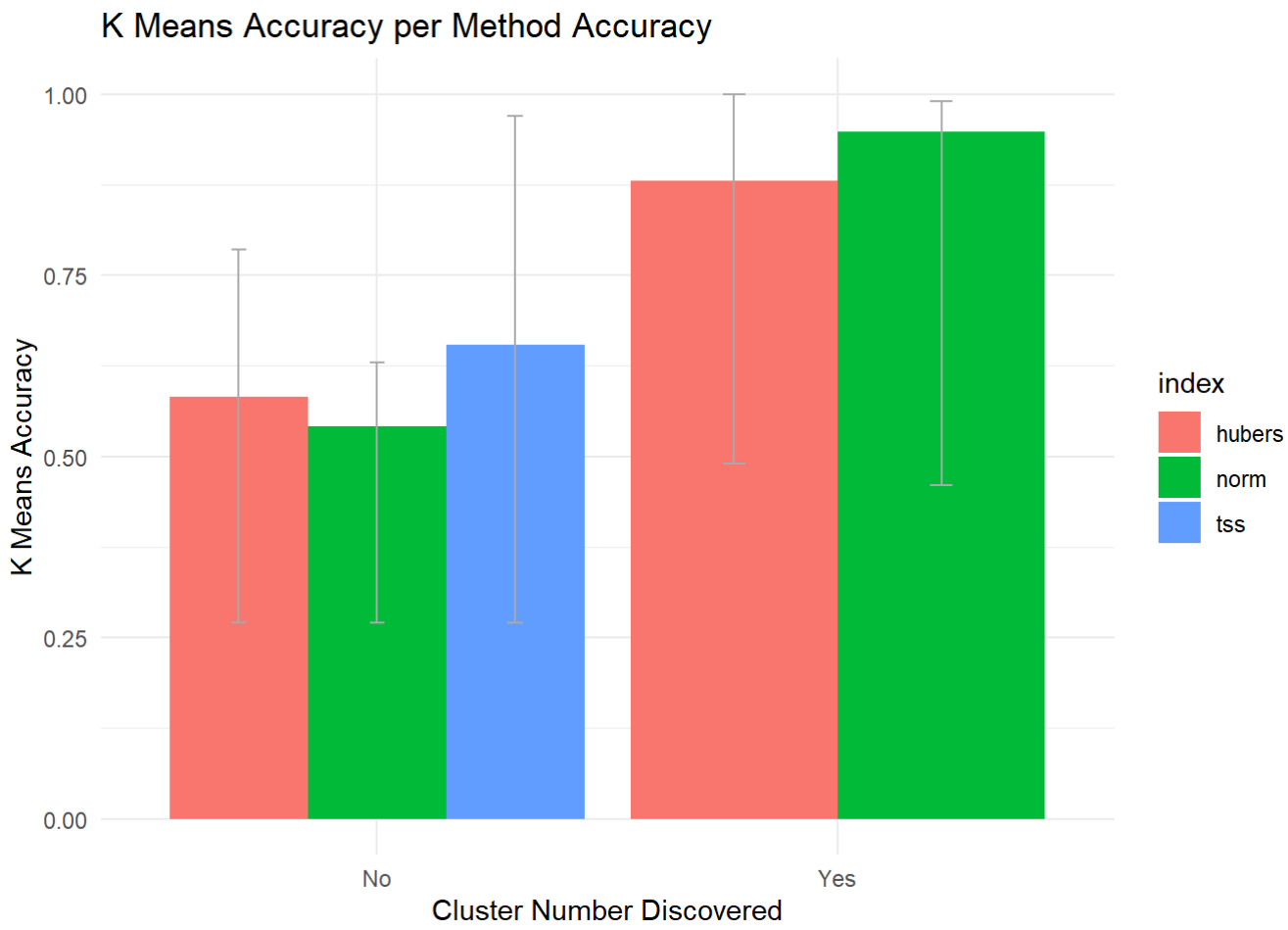
Figure 2 shows how many times each method distribution pairing identified the correct cluster number and did not identify any other cluster number as significant, broken down by seperation and ratio of noise varabibles (max 3). As the ratio of noise variables increasesand the seperation value decreases (top right of each figure) the clustering problem gets harder.

Figure 2 shows huberts gamma statistic performs better than the other 2 metrics and shows a split between random order being better at identifying the harder cluster problems with smaller seperation, and pca better at solving the easier ones. This could be because if there a large seperation in the data already there will also be in the null distribution as it only uses the values that exist.

Figure 3 shows how many times the method, distributer pairing thought there were clusters there (for k = 2 -6). What it shows is within cluster sum of squares unable to desern between clustered and null distributions whatsoever, however the issue with hubers random order and norm min max seem that it is finds clusters when they are not there.



One potential reason for the methods not finding hte correct cluster number is that k-means did a terrible job of identifying the clusters, so we compared the mean matching score between the k means cluster labels and the original cluster. This is shown in figure 4. It appears from this plot that k-means is partly responsible for not being able to identify the correct cluster number



Going Forward

1. Use PCA before K-means with greater number of random starts to improve performance
2. Test more cluster metrics (drop tss)
3. Return metrics on the distributions namely kertosis