On Cluster Reassignment

Abstract

As Andrea has purposed, I manually reassigned outliers based on the mean of each cluster. It seems that manually reassign outliers is not a promising way to optimize our cluster pattern.

Algorithm

In my code, I loop through each outlier. Within this loop, I then loop through all clusters, calculate their mean, and then assign the outlier to the cluster whose mean is closest to the outlier's value. No reassignment happens if the outlier currently belongs to the cluster whose mean is already closest.

Note that the mean of clusters can change with each reassignment. For example, suppose cluster 1.1 previously has [10, 20], then 60 is reassigned to this cluster, the cluster's mean will change from 15 to 30. This feature of my program ensures all reassignment to be more precise.

Since I am playing with one dimensional data, I don't see any other possible algorithm.

Result

As is shown in two csv files, before reassignment we have 468 outliers, but afterwards we have 476. I think this is enough to show that manual reassignment is not very promising in optimizing our cluster pattern.

Explanation

Why does manual reassignment fail? The reason is that we create new outliers while eliminating old ones. For example in cluster 1.1.2.4.2, before reassignment we have 14 outliers, afterwards we have 35. However, it remains unclear how many of these new outliers are previously outliers in other clusters that get reassigned here, and how many of them are previously good data point that become outliers as the cluster is skewed by reassigned data.

Next Step

As for the next step, I think it would be good to try Gaussian Mixture Model proposed by Ali. I believe a new clustering model, being systematic and time-tested, is likely better than our effort without theoretical support. And learning about new algorithm might give us insight on our problem now.