Consensus Clustering of Discrete Data Through Earth Mover's Distance Decision Partitions and Dynamic Hierarchical Threshold Selection

By: Alexander Kiefer & Taufique Hussain

Abstract

In order to provide accurate and confident predictions for classification tasks, particularly group membership, a common technique used by data scientists are consensus functions. Consensus functions are capable of synthesizing several differently classified data points into a single consensus. Through this process, it is possible to increase the confidence and accuracy of a naive classification and, overall, improve results. In our research, we aim to analyze the effectiveness of our formulated consensus function which leverages bipartite graph partitions and the Earth Mover's Distance Algorithm to synthesize multiple data clusterings into a single, highly accurate clustering. In our work, we place emphasis on the analysis of how threshold selection on the local and global level affects the outcome of our consensus function.

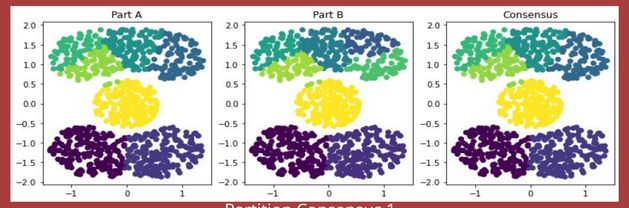
Introduction

The structure of our proposed algorithm utilizes bipartite graphs to hierarchical assess and combine multiple datasets. To begin, any number of labeled inputs are fed into the algorithm. In our testing, we used the artificially generated Cassini dataset available within the R package Clue. We then fed nine versions of the dataset into the algorithm, each being a distinctly labeled set of the same points using K-means clustering from 2 to 10. Once this was done, the algorithm calculates the two most similar input partitions. The two most similar are then prepared for the EMD algorithm by constructing a complete bipartite graph, with edge weights between clusters being calculated through Jaccard similarity. At this point, the algorithm is run through EMD, with flow values being calculated between each cluster in each partition.

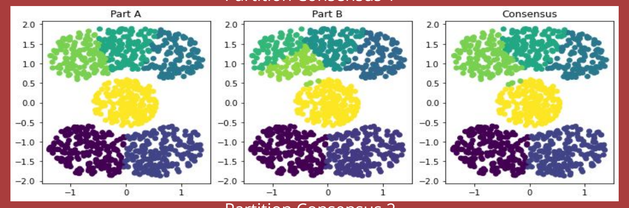
At this point, we reach the experimental portion of our research, which aims to determine whether the threshold, used for determining the minimum flow between partitions, should be determined at the local, or global level of the algorithm. The distinction between local and global, in terms of our algorithm, is whether the threshold should be chosen independently for every two partitions within the hierarchy, or, if it should be assessed as a constant throughout the hierarchy, dependent upon only the final consensus. In addition to where it should be chosen, we research the method by which this threshold should be chosen, as a reliable indicator for clustering quality is essential for the algorithm to function consistently.

Methodology

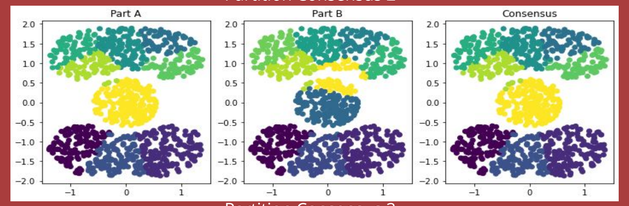
Analysis of Hierarchical Structure for Manually Chosen Global Threshold



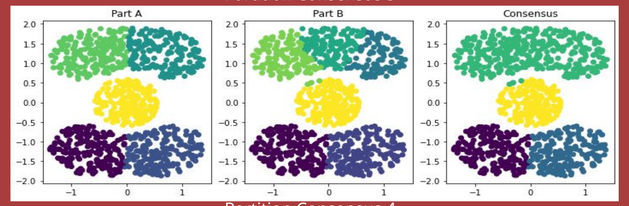
Partition Consensus 1



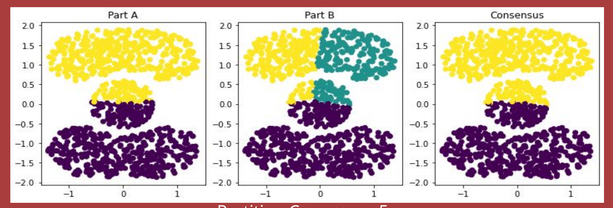
Partition Consensus 2



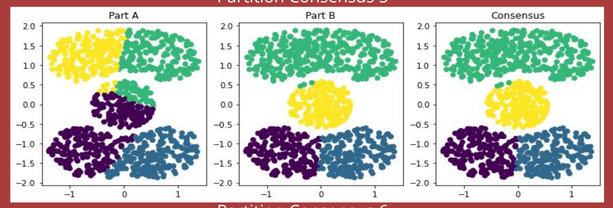
Partition Consensus 3



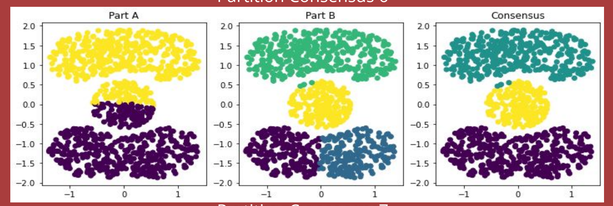
Partition Consensus 4



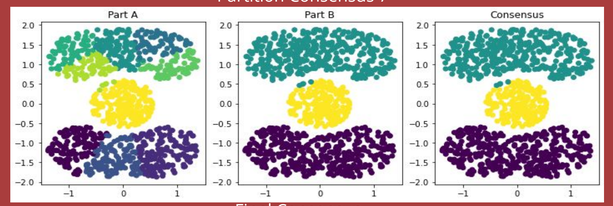
Partition Consensus 5



Partition Consensus 6



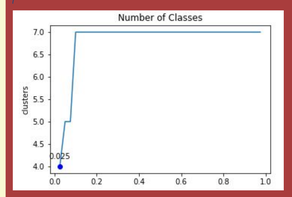
Partition Consensus 7

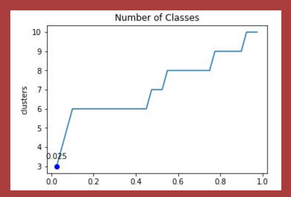


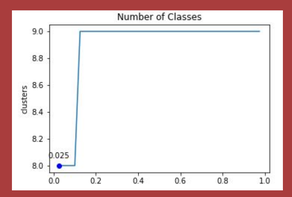
Final Consensus

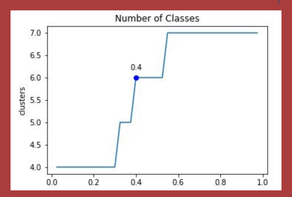
Trend Analysis of Number of Clusters vs Threshold Value for Local and Global Implementations

Local

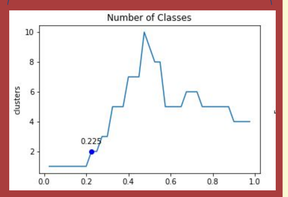








Global



Results

In order to compare and contrast the effectiveness of threshold selection on local and global scales, we conducted two primary tests. First, in order to better understand the effects of threshold selection for both levels, we plotted the hierarchy of bipartite consensuses made through the course of the algorithm. Included in the Methodology section, the (A) and (B) partitions each represent either an input to the algorithm, or a consensus previously reached within the hierarchy, and the consensus represents the output of the algorithm for that level of the hierarchy. With the hierarchy beginning on the top left and ending on the bottom right, we can see just how each level contributes to the overall output. Though we also did this on the local level, we were not able to include all of the graphs. As such, those can be found in the included Github repository.

The second test which we conducted was an analysis of how the number of clusters in consensus function varied with relation to the threshold. This test would be the deciding factor for our results, as seeing an expected trend would give us a scientific basis for choosing one method over the other. With this test, we were able to show that when the threshold is chosen on the local level, the expected trend of lower values producing fewer clusters and higher values producing more is proven to be true. When compared to the threshold selection on the global level, where the number of clusters is inconsistent, and does not follow a general trend, we find that by using a locally defined threshold value, that better control of the output can be achieved.

Conclusion

Overall, the results of our research were promising, as our algorithm performed in the way that we had hoped. With local threshold selection proving to be superior to global selection, our research has continued into an evaluation of various functions for threshold selection, with our primary candidate currently being MANOVA, or Multivariate analysis of variance. Though the results of our tests have, so far, been inconclusive, we hope to modify our algorithm to remove the current tie breaking mechanism of voting, in order to eliminate error from propagating forward. In addition, we hope to further extend our analysis to determine the functionality of our algorithm of datasets to be different data points, with the ultimate goal of producing a consensus function capable of efficiently and effectively computing consensuses for massive, continually growing datasets.

References

https://github.com/alexk101/research_sp21