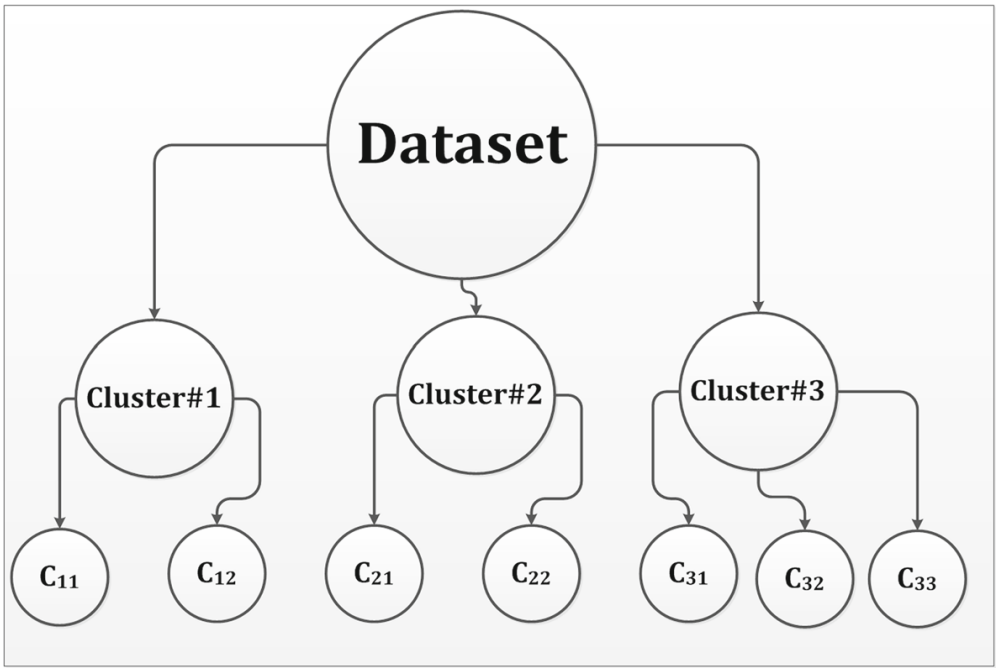
Infrequent pattern mining in smart healthcare environment using data summarization

The goal of summarization is to provide an overview of the data that helps an analyst gain the gist of it and act accordingly.

The problem: Modern data summarization techniques does not capture infrequent patterns in the data because the proportion of these patterns are very low in the parent data compared to the predominantly frequent ones.

Proposed Approach: The algorithm SIPPS proposed in the paper has suggested recursive clustering for data summarization which can be defined formally; to create clusters from a given dataset, and then, each of the individual clusters (also termed as parent cluster) is clustered again to divulge the underlying groups in those clusters. Considering Fig. 3, the clusters $C11, C12, C21, C22, C31, C32, C33$ can be defined as the recursive clusters.



The idea is that, instead of selecting data samples from the parent clusters, it might be more informative to select these samples from the recursive clusters which would contain more finely grained representation of the dataset, hence infrequent patterns which are small can be found.

We can adapt this model for our task by using a stopping criteria, where we stop the clustering if a threshold is reached depending on the size of our dataset or till 2 clusters are found where 1 of them cluster has only 1 instance as was done in the paper.

Evaluation: For evaluating the model, they benchmarked a set of datasets where they tested 5 data summarization techniques and their approach and it showed that SIPPS performed better based on the metric “Infrequent Pattern Loss”.