

Super Partition: fast, flexible, and interpretable large-scale data reduction in R

April 22, 2024

1 Supplemental Material

Example Code

```
library(GEOquery)
library(partition)

# query CAMP gene expression data from GEO
## accession number – GSE22324
camp <- getGEO("GSE22324", GSEMatrix = TRUE)
dim(camp$GSE22324_series_matrix.txt.gz)

# get expression matrix
camp_ge <- as.data.frame(t(exprs(camp$GSE22324_series_matrix.txt.gz)))
dim(camp_ge)

# information loss criterion (ILC) values
ILClist <- seq(from=0, to=1, by=0.05)

# for each ILC
for (i in 1:length(ILClist)) {
  start.time <- Sys.time()

  # partition data at appropriate ILC
  camp_part <- super_partition(full_data = camp_ge,
                              threshold = ILClist[i],
                              cluster_size = 4000)

  # print time and number of features
  end.time <- Sys.time()
  diff <- difftime(end.time, start.time, units='mins')
  print(paste0("ILC = ", ILClist[i], " number of features: ",
              dim(camp_part$reduced_data)[[2]]))
  print(diff)
}
```