NBHF Clusters: Streamlined Results

Jackson Vanfleet-Brown

2024-02-25

## Intro

We want to identify clusters within our training set before we make our classifier learn from it. There are two objectives for unsupervised clustering:

1. Look across species to identify whether species classes form separate clusters ([Section 1](#sec-species)).
   * The existence of clusters suggests that there are meaningful differences between classes that the classifier can be trained to recognize.
2. Look within species classes to assess variability among events ([Section 2](#sec-events)).
   * The existence of clusters within an individual species class may indicate that there are outlying events with anomalous features that should be excluded from the training set.

## Objective 1

### Method

set.seed(123)  
  
# slice sample of 200 clicks from each species  
samp <- nbhf\_clicks %>%   
 group\_by(species) %>%  
 slice\_sample(n=200) %>%  
 ungroup()  
  
samp\_rm <- samp %>%   
 # drop metadata  
 select(-c(UID:noiseLevel, BinaryFile, eventLabel,detectorName, db)) %>%   
 # drop variables to avoid creating artifacts in the cluster plot.  
 select(species, eventId, duration:peak, Q\_10dB:centerkHz\_3dB)  
  
# calculate Euclidean distances  
dist <- samp\_rm %>%  
 select(-c(species, eventId)) %>%  
 mutate(id = 1:n()) %>%  
 column\_to\_rownames("id") %>%  
 scale() %>%  
 dist(method="euclidean")  
  
cl <- densityClust(dist)  
# set rho and delta values  
cl <- findClusters(cl, rho=25, delta=2)

Using the above method, the density clustering algorithm formed the clusters shown in [Figure 1](#fig-density-clust). The counts of each species in each of the resulting clusters is given in [Table 1](#tbl-clust-assn). The MDS plot is shown with the points colored by species in [Figure 2](#fig-mds-species).

|  |
| --- |
| Figure 1: Density clusters with Four clusters formed with ρ=25 and δ=2 |

|  |
| --- |
| Table 1: Table of cluster assignments  1 2 3 4  ks 2 4 3 191  pd 170 10 0 20  pp 59 137 0 4 |

|  |
| --- |
| Figure 2: MDS plot showing distances between clicks in the training set, colored by species |

### Discussion

* For all species, we see that clicks cluster predominately in a single cluster, with little overlap between different species classes.
* The MDS plot similarly shows that clicks separate from one another on the basis of species class.
* From [Table 1](#tbl-clust-assn) we see that cluster 3 is very small, containing just three Kogia clicks.
  + [Table 2](#tbl-clust-ks) shows that the entirety of cluster 3 is formed from a few anomalous clicks in *Kogia* event identified as PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE4.
  + That same event, which happens to be the largest *Kogia* event in the entire training set, has the overwhelming majority of it’s clicks in cluster 4.

|  |
| --- |
| Table 2: Table of cluster assignments for all *Kogia* events  1 2 3 4  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE1 0 0 0 4  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE2 0 0 0 9  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE3 0 0 0 9  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE4 0 0 0 12  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE5 0 0 0 34  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE6 0 0 0 3  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE7 0 0 0 11  PG2\_02\_09\_CCES\_022\_Ksp - Copy.OE8 1 0 0 9  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE1 0 0 0 7  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE2 0 0 0 5  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE3 0 0 0 8  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE4 1 0 3 46  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE5 0 0 0 5  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE6 0 2 0 18  PG2\_02\_09\_CCES\_023\_Ksp - Copy.OE7 0 2 0 11 |

## Objective 2

We will now subset the training data by species and then re perform density clustering to identify anomalous events.

### Method

set.seed(123)  
  
samp2 <- nbhf\_clicks %>%  
 nest(data=-species) %>%  
 # determine the median event n for each species  
 mutate(median\_ev\_n=map\_dbl(data,\(d) d %>% count(eventId) %>% pull(n) %>% median())) %>%  
 # for events with n greater than the median, slice a sample in size equal to the median  
 # since n varies across two orders of magnitude, intended to balance representation in the plot  
 mutate(samp=map2(data, median\_ev\_n, \(d, s) d %>% group\_by(eventId) %>% slice\_sample(n=s) %>% ungroup())) %>%  
 # select variables of interest  
 mutate(samp=map(samp, \(s) select(s, eventId, duration:peak, Q\_10dB:centerkHz\_3dB))) %>%  
 mutate(samp=map(samp, \(s) s %>% mutate(id=1:n()) %>% drop\_na())) %>%  
 select(species, samp) %>%   
 unnest(samp)  
  
  
sp <- c("ks", "pd", "pp")  
# subset data by species  
clicks <- lapply(sp, \(x) filter(samp2, species==x))  
# create distance matrices.  
dist2 <- lapply(clicks, \(c) c %>% select(-c(species, eventId)) %>% column\_to\_rownames("id") %>% scale() %>% dist())  
cl2 <- lapply(dist2, densityClust)  
# Perform density clustering. Static values chosen for rho and delta.  
# This decision does not seem to be critical, because the algorithm strongly favors a single cluster for each species.  
cl2 <- lapply(cl2, findClusters, delta = 8, rho = 5)

[Figure 3](#fig-event-clusters) shows the resulting density cluster plots and [Figure 4](#fig-mds-events) shows the plots with the points colored by event.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | (a) *Kogia* |  |  | | --- | | (b) Dall’s porpoise |  |  | | --- | | (c) Harbor porpoise |   Figure 3: Click clusters for each species class |

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | (a) *Kogia* |  |  | | --- | | (b) Dall’s porpoise |  |  | | --- | | (c) Harbor porpoise |   Figure 4: MDS plot showing distances between clicks, colored by event. Legend is hidden due to large number of events in each species class. |

### Discussion

* The density clustering algorithm appears to strongly favors a single cluster in each species, suggesting that there are no outlying events.
* When points are colored by event, variation among events is more evident. This variation does not appear to be strong enough to manifest as more than one density-based cluster.