BFL_synthetic_PHMRC_example

Yu (Zoey) Zhu

2025-05-09

```r
library(rstan)
library(caret)
library(LCVA)
```

This supplementary R Markdown file provides example code for fitting Local and Bayesian Federated Learning (BFL) models under three different label shift settings: (1) No within-target label shift, (2) Mild within-target label shift, and (3) Severe within-target label shift. The goal is to evaluate model performance on a given target domain (e.g., "AP") by computing both the Top Cause Accuracy and the Cause-Specific Mortality Fraction (CSMF) Accuracy.

The "execute_balance_local_fit" and "execute_unbalance_local_fit" functions generate simulated data for each domain, fit the local LCVA models for each domain, and extract the corresponding posterior phi for subsequent BFL models. These functions also report the local model's CSMF and top-cause accuracy for the target site.

The "run_BFL_balance_model" and "run_BFL_unbalance_model" functions fit various BFL models and provide both CSMF and top-cause accuracy predictions for the target site. Additionally, the outputs also include CSMF estimates and individual-level cause predictions for the unlabeled samples.

Setup

We consider the target domain is "AP" and have 20% labeled data in the target domain (label shift case (1) & (2)).
We apply K = 5 for the latent class when fitting LCVA base models.

```r
test_site_balance <- "AP"
test_site_unbalance <- "AP"
sites <- c("Mexico", "AP", "Bohol", "Dar", "Pemba", "UP")
K <- 5
miss_prop <- 0.8
model_types_balance <- c("domain", "partial", "base", "mix")
model_types_unbalance <- c("domain", "partial", "mix")
```

Case I: No within-target label shift

We evaluate the local models (local-self and local-avg), and BFL models (domain, partial, base, mix) in the 'no within-target label shift' case.

```
source("execute_balance_local_fit.R")
source("run_BFL_model.R")
source("execute_balance_domain.R")
source("execute_balance_partial.R")
source("execute_balance_base.R")
source("execute_balance_mix.R")
```

Local Fit

```
local_res_balance <- execute_balance_local_fit(
  test_site = test_site_balance,
  sites = sites,
  K = K,
  miss_prop = miss_prop
)

posterior_phi_full <- local_res_balance$posterior_phi_full
sim_data_filtered_list <- local_res_balance$sim_data_filtered_list
LCVA_local_model_test_obs_fit <- local_res_balance$LCVA_local_model_test_obs_fit
```

```
# CSMF ACC
cat("Case I : Local-self :", local_res_balance$csmf_acc_local[test_site_balance], "\n")
```

```
## Case I : Local-self : 0.6148636
```

```
cat("Case I : Local-avg :",
    mean(local_res_balance$csmf_acc_local[names(local_res_balance$csmf_acc_local) != test_site_balance]
```

```
## Case I : Local-avg : 0.5925055
```

```
# Top Cause ACC
cat("Case I : Local-self :", local_res_balance$acc_local[test_site_balance], "\n")
```

```
## Case I : Local-self : 0.2485921
```

```
cat("Case I : Local-avg :",
    mean(local_res_balance$acc_local[names(local_res_balance$acc_local) != test_site_balance]), "\n")
```

```
## Case I : Local-avg : 0.2366854
```

BFL Fit

```
BFL_results_balance <- list()

for (model_type in model_types_balance) {
  BFL_results_balance[[model_type]] <- run_BFL_balance_model(
    model_type = model_type,
```

```r
    test_site = test_site_balance,
    sites = sites,
    sim_data_filtered_list = sim_data_filtered_list,
    posterior_phi_full = posterior_phi_full,
    LCVA_local_model_test_obs_fit = LCVA_local_model_test_obs_fit
  )
}
```

```
## Stacking 3 chains, with 152 data points and 1500 posterior draws;
##  using stan optimizer, max iterations = 1e+05
##
##  Total elapsed time for approximate LOO and stacking = 16.87 s
## [1] "Number of posterior draws from each chain:"
## w[1] w[2] w[3]
## 1222 1458 1320
```

```r
# CSMF ACC
for (model_type in model_types_balance) {
   cat("Case I : BFL-", model_type, ":", BFL_results_balance[[model_type]]$csmf_acc, "\n")
}
```

```
## Case I : BFL- domain : 0.6893699
## Case I : BFL- partial : 0.7107892
## Case I : BFL- base : 0.5910754
## Case I : BFL- mix : 0.6998775
```

```r
# Top Cause ACC
for (model_type in model_types_balance) {
  cat("Case I : BFL-", model_type, ":", BFL_results_balance[[model_type]]$acc, "\n")
}
```

```
## Case I : BFL- domain : 0.3459372
## Case I : BFL- partial : 0.3419147
## Case I : BFL- base : 0.3082368
## Case I : BFL- mix : 0.3419147
```

```r
# example of CSMF estimation from BFL-domain
BFL_results_balance[["domain"]]$csmf
```

```
##                       Cirrhosis                        Epilepsy
##                     0.035859600                     0.005989274
##                       Pneumonia                            COPD
##                     0.077476216                     0.008486720
##     Acute Myocardial Infarction                           Fires
##                     0.103730125                     0.022662488
##                   Renal Failure                            AIDS
##                     0.111075462                     0.162055543
##                     Lung Cancer                        Maternal
##                     0.001751973                     0.063485187
##                        Drowning   Other Cardiovascular Diseases
##                     0.032734139                     0.014222926
## Other Non-communicable Diseases                           Falls
##                     0.018279802                     0.015194428
##                          Stroke                    Road Traffic
##                     0.063322962                     0.016319234
##         Bite of Venomous Animal                        Diabetes
##                     0.007956630                     0.031037776
##       Other Infectious Diseases                              TB
##                     0.092282992                     0.019473165
##                         Suicide                  Other Injuries
##                     0.009946256                     0.015708609
##                 Cervical Cancer                         Malaria
##                     0.001363701                     0.006491573
##                          Asthma              Diarrhea/Dysentery
##                     0.003211193                     0.012320153
##               Colorectal Cancer                        Homicide
##                     0.001450121                     0.024795011
##                   Breast Cancer              Leukemia/Lymphomas
##                     0.001295666                     0.001278824
##                      Poisonings                 Prostate Cancer
##                     0.014332290                     0.001281151
##               Esophageal Cancer                  Stomach Cancer
##                     0.001267687                     0.001861126
```

```
# example of individual prediction for the unlabeled sample from BFL-domain (first 10 samples)
BFL_results_balance[["domain"]]$cause_pred[1:10]
```

```
##  [1] "Pneumonia"                     "Pneumonia"
##  [3] "Other Cardiovascular Diseases" "Pneumonia"
##  [5] "TB"                            "Maternal"
##  [7] "Diarrhea/Dysentery"            "Other Infectious Diseases"
##  [9] "Maternal"                      "Maternal"
```

Case II: Mild within-target label shift

We evaluate the local models (local-self and local-avg), and BFL models (domain, partial, mix) in the 'mild within-target label shift' case.

```
source("execute_unbalance_local_fit.R")
source("run_BFL_model.R")
source("execute_unbalance_domain.R")
source("execute_unbalance_partial.R")
source("execute_unbalance_mix.R")
```

```
unbalanced_cases <- c("MILD", "SEVERE")
```

Local Fit

```
case = unbalanced_cases[1]
cat("Running Label Shift Case:", case, "\n")
```

```
## Running Label Shift Case: MILD
```

```r
local_res_unbalance <- execute_unbalance_local_fit(
    test_site = test_site_unbalance,
    sites = sites,
    K = K,
    miss_prop = miss_prop,
    unbalanced_case = case
  )
```

```r
# CSMF ACC
cat("Case II : Local-self :", local_res_unbalance$csmf_acc_local[test_site_unbalance], "\n")
```

```
## Case II : Local-self : 0.4790563
```

```r
cat("Case II : Local-avg :",
    mean(local_res_unbalance$csmf_acc_local[names(local_res_unbalance$csmf_acc_local) != test_site_unba
```

```
## Case II : Local-avg : 0.5441316
```

```r
# Top Cause ACC
cat("Case II : Local-self :", local_res_unbalance$acc_local[test_site_unbalance], "\n")
```

```
## Case II : Local-self : 0.2886179
```

```r
cat("Case II : Local-avg :",
    mean(local_res_unbalance$acc_local[names(local_res_unbalance$acc_local) != test_site_unbalance]), "
```

```
## Case II : Local-avg : 0.2260163
```

BFL Fit

```r
  BFL_results_unbalance <- list()

  for (model_type in model_types_unbalance) {
    BFL_results_unbalance[[model_type]] <- run_BFL_unbalance_model(
      model_type = model_type,
      test_site = test_site_unbalance,
      sites = sites,
      sim_data_filtered_list = local_res_unbalance$sim_data_filtered_list,
      sim_data_target_domain_list = local_res_unbalance$sim_data_target_domain_list,
      posterior_phi_full = local_res_unbalance$posterior_phi_full,
      LCVA_local_model_test_obs_fit = local_res_unbalance$LCVA_local_model_test_obs_fit
    )
  }
```

```
## Stacking 3 chains, with 652 data points and 1500 posterior draws;
##  using stan optimizer, max iterations = 1e+05
##
```

```
##  Total elapsed time for approximate LOO and stacking = 18 s
## [1] "Number of posterior draws from each chain:"
## w[1] w[2] w[3]
##  535 1130 2335
```

```
# CSMF ACC
for (model_type in model_types_unbalance) {
   cat("Case II : BFL-", model_type, ":", BFL_results_unbalance[[model_type]]$csmf_acc, "\n")
}
```

```
## Case II : BFL- domain : 0.5892829
## Case II : BFL- partial : 0.5029196
## Case II : BFL- mix : 0.5232224
```

```
# Top Cause ACC
for (model_type in model_types_unbalance) {
  cat("Case II : BFL-", model_type, ":", BFL_results_unbalance[[model_type]]$acc, "\n")
}
```

```
## Case II : BFL- domain : 0.296748
## Case II : BFL- partial : 0.2804878
## Case II : BFL- mix : 0.2764228
```

Case III: Severe within-target label shift

We evaluate the local models (local-self and local-avg), and BFL models (domain, partial, mix) in the 'severe within-target label shift' case.

Local Fit

```
case = unbalanced_cases[2]
cat("Running Label Shift Case:", case, "\n")
```

```
## Running Label Shift Case: SEVERE
```

```
local_res_unbalance <- execute_unbalance_local_fit(
    test_site = test_site_unbalance,
    sites = sites,
    K = K,
    miss_prop = miss_prop,
    unbalanced_case = case
  )
```

```
# CSMF ACC
cat("Case III : Local-self :", local_res_unbalance$csmf_acc_local[test_site_unbalance], "\n")
```

```
## Case III : Local-self : 0.04261776
```

```r
cat("Case III : Local-avg :",
    mean(local_res_unbalance$csmf_acc_local[names(local_res_unbalance$csmf_acc_local) != test_site_unba
```

```
## Case III : Local-avg : 0.3080023
```

```r
# Top Cause ACC
cat("Case III : Local-self :", local_res_unbalance$acc_local[test_site_unbalance], "\n")
```

```
## Case III : Local-self : 0.008842444
```

```r
cat("Case III : Local-avg :",
    mean(local_res_unbalance$acc_local[names(local_res_unbalance$acc_local) != test_site_unbalance]), "
```

```
## Case III : Local-avg : 0.1729904
```

BFL Fit

```r
  BFL_results_unbalance <- list()

  for (model_type in model_types_unbalance) {
    BFL_results_unbalance[[model_type]] <- run_BFL_unbalance_model(
      model_type = model_type,
      test_site = test_site_unbalance,
      sites = sites,
      sim_data_filtered_list = local_res_unbalance$sim_data_filtered_list,
      sim_data_target_domain_list = local_res_unbalance$sim_data_target_domain_list,
      posterior_phi_full = local_res_unbalance$posterior_phi_full,
      LCVA_local_model_test_obs_fit = local_res_unbalance$LCVA_local_model_test_obs_fit
    )
  }
```

```
## Stacking 3 chains, with 157 data points and 1500 posterior draws;
##  using stan optimizer, max iterations = 1e+05
##
##  Total elapsed time for approximate LOO and stacking = 16.56 s
## [1] "Number of posterior draws from each chain:"
## w[1] w[2] w[3]
## 1620  670 1710
```

```r
# CSMF ACC
for (model_type in model_types_unbalance) {
   cat("Case III : BFL-", model_type, ":", BFL_results_unbalance[[model_type]]$csmf_acc, "\n")
}
```

```
## Case III : BFL- domain : 0.6026971
## Case III : BFL- partial : 0.614353
## Case III : BFL- mix : 0.6067743
```

```r
# Top Cause ACC
for (model_type in model_types_unbalance) {
  cat("Case III : BFL-", model_type, ":", BFL_results_unbalance[[model_type]]$acc, "\n")
}
```

```
## Case III : BFL- domain : 0.3223473
## Case III : BFL- partial : 0.3118971
## Case III : BFL- mix : 0.3191318
```